Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Bc. Petra Exnarová

Postupy pro detekci změny v některých speciálních regresních modelech

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Prof. RNDr. Marie Hušková, DrSc.

Studijní program: Matematika

Studijní obor: Pravděpodobnost, matematická statistika a ekonometrie — Matematická statistika

Praha 2013

Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS

Bc. Petra Exnarová

Detection of change in some special regression models

Department of Probability and Statistics

Supervisor of the master thesis: Prof. RNDr. Marie Hušková, DrSc.

Study program: Mathematics

Specialization: Probability, Mathematical Statistics and Econometrics — Mathematical Statistics

Prague 2013

Na tomto místě bych ráda srdečně poděkovala prof. RNDr. Marii Huškové, DrSc. za laskavé vedení a cenné rady, které mi poskytla při vytváření této práce, a především za trpělivost a ochotu při udílení konzultací.

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 12.4.2013 Petra Exnarová

Název práce: Postupy pro detekci změny v některých speciálních regresních modelech
Autor: Bc. Petra Exnarová
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Prof. RNDr. Marie Hušková, DrSc.

Abstrakt: Předložená práce se zabývá detekcí změny ve třech speciálních případech. Prvním z nich je případ spojité změny v lineárním modelu (tzv. broken-line model), zbývající dva se zabývají změnou parametru v diskrétním rozdělení — nejdříve je probrán jednodušší případ Bernoulliho rozdělení, který je poté rozšířen na případ multinomického rozdělení. Pro všechny uvedené případy jsou popsány postupy pro známý i neznámý bod změny. Vedle aproximace pomocí limitních vět jsou v této práci popsány také možnosti aproximace pomocí metody zvané bootstrap a permutačního testu pro všechny uvedené případy. Porovnání kritických hodnot získaných různými přístupy a malá studie síly testů jsou realizovány pomocí simulací.

Klíčová slova: detekce změny, broken-line model, diskrétní rozdělení, bootstrap, permutační test

Title: Detection of change in some special regression models
Author: Bc. Petra Exnarová
Department: Department of Probability and Mathematical Statistics
Supervisor: Prof. RNDr. Marie Hušková, DrSc.

Abstract: Presented thesis deals with testing of change in three special cases of change-point analysis. First of them is case of continuous change in linear regression (so-called broken-line model), the other two are related to change in parameters of discrete value distributions — simple case of Bernoulli distributed variables is studied first and then the approach is generalized for case of Multinomial distribution. Both situations of known and unknown change point are described for all three cases. Beside approximation by using limit theorems, the bootstrap method and permutation test are described for all studied cases as well. The comparison of critical values gained by different approaches for the particular tests and small power analysis is done using simulations.

Keywords: change-point analysis, broken-line model, discrete distribution, bootstrap, permutation test

Contents

Used notation

- $\lfloor x \rfloor$ ... floor of x (the largest integer not greater than x)
- $\lceil x \rceil$ ... ceiling of x (the smallest integer not less than x)
- $\boldsymbol{x}^T$, $\boldsymbol{X}^T$ ... transpose of a vector $\boldsymbol{x}$ or matrix $\boldsymbol{X}$
- $\boldsymbol{I}$ ... identity matrix
- $\mathrm{Diag}\{a_{11}, \ldots, a_{nn}\}$ ... diagonal matrix with elements $a_{11}, \ldots, a_{nn}$ on the main diagonal
- $\mathsf{Tr}\, \boldsymbol{A}$ ... trace of a matrix $\boldsymbol{A}$
- $I\,[X = x]$ ... indicator of the set $X = x$ (equals to 1 if $X = x$ and 0 if $X \neq x$)
- $\mathbb{N}$ ... set of all positive integers
- $\mathbb{R}$ ... set of all real numbers
- $\approx$ ... approximation

Introduction

Detection of change at unknown point and estimating the location of the change are common statistical problems known as the change-point problem. A vast literature on this theme for different situations is available. The simplest case of change-point analysis is a change in a parameter of some known distribution. We can imagine this, for example, as a manufacturing process generating some components of concrete size with controlled variance. It can happen that due to a failure, the distribution of the component size changes and we would like to know when it happened. The testing of change and the estimation of the change point for this simple case can be found for example in Antoch et al. [2000]. A broad variation of change-point situations in linear regression can be found in literature. The change can happen in intercept or slope (or in both parameters), the starting parameters can be known or unknown, the change can be continuous or discontinuous, there can be only one explanatory variable or there can be more of them (some of them can be nuisance), the distribution of random errors can be assumed to be normal or not etc. The motivation for these types of change-point analysis is broad as well — meteorological or hydrological data are very frequent, for example. The testing of change and estimation of the change point for some special cases can be found, for example, in Hušková and Antoch [2003], Julious [2001], Jarušková [1998], Antoch et al. [2000]. Medical data are common motivation for change-point analysis in generalized linear models — for example modeling the dependency of risk of some disease on patient's age can show a sharp increase after reaching certain age. The testing of change in generalized linear models can be found for example in Antoch et al. [2004], the estimation of the change point can be found, for example, in Zhou and Liang [2008]. Many other situations related to change-point problem are studied in literature as well, for example change in nonlinear regression [Ciuperca, 2004, 2009] and in hazard regression model [Dupuy, 2006].

In this thesis we deal with testing of change in three special cases. First of them is case of continuous change in linear regression (so-called broken-line model), the other two are related to change in discrete value distributions parameters — simple case of Bernoulli distributed variables is studied first and then the approach is generalized for case of Multinomial distribution.

First, some basic theoretical background as a base for the later chapters is introduced in Chapter 1. In Chapter 2 one can find the overview of broken-line model testing of change – three simple cases of the random error distribution with known change point, one possible approach for unknown change point and algorithms for bootstrap and permutation test approximations are introduced. The change in discrete distribution is the main theme of Chapter 3. First part of the chapter deals with Bernoulli distributed variables — two different test statistics for known change point and one test statistic for unknown change point are studied. Second part of the chapter is dedicated to extension of the results for Multinomial distribution. Chapters 4 and 5 consist of simulation studies for both broken-line and discrete distribution cases.

Some of the theorems used in Chapters 2 and 3 can be found in Appendix A. The statistical software R version 2.15.2 [R Development Core Team, 2010] has been used for simulations. The important parts of the code can be found in Appendix B, whole code is available on the attached CD.

Chapter 1

Theoretical background

Before we start to deal with the main theme of this thesis, we briefly introduce some basic terminology and facts that are going to be useful in future chapters.

1.1 Linear model and estimation methods

Let $\boldsymbol{X}$ be a $n \times k$ matrix of constants, $\boldsymbol{\beta}$ be a k-dimensional vector of unknown parameters and $\boldsymbol{Y}$ be n-dimensional random vector. The vector $\boldsymbol{Y}$ is said to follow the *linear model*, if $\mathsf{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{V} = \operatorname{var} \boldsymbol{Y}$ exists and does not depend on $\boldsymbol{\beta}$. Let $k < n$ and there exists a $k \times n$ matrix $\boldsymbol{U}$ such that $\boldsymbol{b} = \boldsymbol{U}\boldsymbol{Y}$, then the k-dimensional vector $\boldsymbol{b}$ is called *linear estimator* of the vector of parameters $\boldsymbol{\beta}$. The noise element of the above model $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{\beta}\,\boldsymbol{X}$ is a vector with $\mathsf{E}\,\boldsymbol{\varepsilon} = \boldsymbol{0}$ and $\operatorname{var} \boldsymbol{\varepsilon} = \sigma^2 \boldsymbol{I}$, where $\sigma^2 > 0$. The elements of $\boldsymbol{\varepsilon}$ are called *random errors*.

The estimator $\boldsymbol{b}$ is said to be *unbiased* if $\mathsf{E}\boldsymbol{b} = \boldsymbol{\beta}$.

The estimator $\boldsymbol{b}$ is said to be *consistent* if $\boldsymbol{b} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}$ for $n \to \infty$.

The estimation of parameters by minimizing $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ is called *method of least squares*. The least squares estimator of the vector of parameters $\boldsymbol{\beta}$ in the above model can be expressed as $\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ (proof can be found, for example, in Anděl [2005, page 81]).

1.2 Hypothesis testing

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a random vector with distribution depending on a parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta$ where Θ is a set of all possible values of $\boldsymbol{\theta}$ called *parameter space*. If we expect that true value of parameter $\boldsymbol{\theta}$ belong to Θ_0, which is a subset of parameter space Θ, and we want to test this hypothesis, the statement $\boldsymbol{\theta} \in \Theta_0$ is called *null hypothesis* and we denote $H_0 : \boldsymbol{\theta} \in \Theta_0$. Denote $\Theta_1 = \Theta \setminus \Theta_0$, then option $\boldsymbol{\theta} \in \Theta_1$ is called *alternative hypothesis* and we denote $H_1 : \boldsymbol{\theta} \in \Theta_1$. The decision about rejecting or accepting the null hypothesis is based on the vector $\boldsymbol{X}$. We choose a set W such that we reject the null hypothesis if $\boldsymbol{X} \in W$ and do not reject the null hypothesis if $\boldsymbol{X} \notin W$. The set W is called *critical region*. The situation when we reject the null hypothesis while the null hypothesis is true is called *Type I error*. On the other hand, the situation when we do not reject the null hypothesis and the null hypothesis is not true is called *Type II error.*

We choose the critical region such that the Type I error is less or equal to a small number α (usually we choose $\alpha = 5\%$). The number $\sup_{\boldsymbol{\theta} \in \Theta_0} \mathsf{P}(\boldsymbol{X} \in W)$ is called *level of significance.* The *power function* of the test is defined as $\pi(\boldsymbol{\theta}) = \mathsf{P}_{H_1}(\boldsymbol{X} \in W) = \mathsf{P}(\boldsymbol{X} \in W | \boldsymbol{\theta} \in \Theta_1)$ (probability of rejecting the null hypothesis if the alternative hypothesis is true).

1.3 Matrices

The number of linearly independent rows or columns in a $m \times n$ matrix $\boldsymbol{A}$ is called *rank* of the matrix $\boldsymbol{A}$.

A $n \times n$ square matrix $\boldsymbol{A}$ is said to be *nonsingular* if the rank of the matrix is equal to n.

A square matrix $\boldsymbol{A}$ is said to be *idempotent* if $\boldsymbol{A}^2 = \boldsymbol{A}$.

A $n \times n$ square matrix $\boldsymbol{A}$ is said to be *positive definite* if it is symmetric and $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0$ for any n-dimensional vector $\boldsymbol{x}$.

A $n \times n$ square matrix $\boldsymbol{A}$ is said to be *positive semi-definite* if it is symmetric and $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0$ for any n-dimensional vector $\boldsymbol{x}$.

Let $\boldsymbol{A}$ be a $m \times n$ matrix, then a $n \times m$ matrix $\boldsymbol{A}^-$ is called *pseudoinverse* of the matrix $\boldsymbol{A}$ if

$$\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}.$$

1.4 Convergence

Let $X, X_1, X_2, \ldots$ be random variables defined on the same probability space. The sequence of random variables $X_1, X_2, \ldots$ is said to *converge in probability* towards X if

$$\lim_{n \to \infty} \mathsf{P}(|X_n - X| > \varepsilon) = 0 \quad \text{for } \forall \varepsilon > 0.$$

We denote this $X_n \xrightarrow{\mathsf{P}} X$.

The sequence of random variables $X_1, X_2, \ldots$ is said to *converge almost surely* towards X if

$$\mathsf{P}\left(\lim_{n \to \infty} X_n = X\right) = 1.$$

We denote this $X_n \xrightarrow{a.s.} X$.

Denote F_n the distribution function of a random variable X_n, $n \in \mathbb{N}$, and F the distribution function of a random variable X. The sequence $\{F_1, F_2, \ldots\}$ is said to *converge weakly* towards F if

$$\lim_{n \to \infty} F_n(x) = F(x), \quad \text{for } \forall x \text{ at which } F \text{ is continuous.}$$

If the distribution functions of $\{X_1, X_2, \ldots\}$ converge weakly towards the distribution function of X, the random variables $X_1, X_2, \ldots$ are said to *converge in distribution* towards the variable X. We denote this $X_n \xrightarrow{d} X$.

1.5 Wiener process and Brownian bridge

A continuous-time stochastic process $\{W(t),\, t \geq 0\}$ with these properties:

1. W(0) = 0 almost surely
2. The increments $W(t_1), W(t_2) - W(t_1), W(t_3) - W(t_2), \ldots, W(t_n) - W(t_{n-1})$, where $0 \leq t_1 < t_2 < \cdots < t_n$, are independent random variables
3. The increments $W(s) - W(t)$ are normally distributed with zero expectation and variance $\sigma^2\,(s-t)$, $\sigma^2 > 0$ for any s and t such that $0 \leq t < s$

is called *Wiener process.*

A Wiener process $\{B(t),\, t \in (0,1)\}$ with property $B(1) = 0$ is called *Brownian bridge.*

1.6 Bootstrap and Permutation test

Bootstrap

The method called *bootstrap* was first introduced by Bradley Efron in Efron [1979]. This method is one of the so-called "computer intensive methods" and among others it can be used for approximation of the test statistics distribution. Like the title of the group of methods says, the usage of this method is highly connected to the computers because it's nearly impossible to compute it by hand already for very small amount of data. The general algorithm of the bootstrap for independent identically distributed observations is following (according to Pastorková [1994]):

1. construct the empirical distribution function from the original random sample $X_1, \ldots, X_N$:
$$G_N(x) = \frac{1}{N} \sum_{i=1}^{N} I(X_i \leq x)$$
2. sample N values from the empirical distribution function $G_N(x)$ and denote $X_1^*, \ldots, X_N^*$; we will call this *bootstrap sample*
3. construct the empirical distribution function from this bootstrap sample:
$$G_N^*(x) = \frac{1}{N} \sum_{i=1}^{N} I(X_i^* \leq x)$$

We described how to obtain one bootstrap sample. As the bootstrap sampling is based on sampling with replacement, there is N^N possible bootstrap samples of size N. Each of these samples has the same probability to be chosen ($\frac{1}{N^N}$), hence we can compute the exact distribution only for very small sample sizes. For the larger amount of data we need to approximate it. The usual option is to repeat the bootstrap algorithm B times, where $N < B < N^N$. From this procedure we obtain B estimates of the distribution function, make a histogram from them and estimate what we need (whole distribution, some parameter, test statistic or quantile). It can be shown that under the null hypothesis the bootstrap is working in case of approximation of a test statistic when particular conditions on the test statistic hold [see Davison and Hinkley, 1997]. All the test statistics described in this thesis satisfy the conditions, but verifying this fact is not part of this thesis.

The principle of the bootstrap method is based on fact that the empirical distribution function of the original sample $G_N(x)$ and the empirical distribution function of the bootstrap sample are close to each other — this follows from advanced results published, for example, in Davison and Hinkley [1997]. In next chapters we use bootstrap for approximating critical values of different tests with different test statistics. It can be shown that in such case bootstrap works only if the sampling is done under the null hypothesis. However, it works even if the original random sample does not satisfy the null hypothesis. More about bootstrap and its use can be found, for example, in the above mentioned Davison and Hinkley [1997].

Permutation test

The first use of permutation test was registered already back in 1930s — the permutation tests for two-sample problems, simple linear regression and block designs were described in Pitman [1937a,b, 1938]. The permutation test is based on relabeling the observations and computing all the possible values of the particular test statistic (for all possible permutations of labels). This provides test on exact significance level that we need. However, there is $N!$ different permutations (where N is the number of observations), so computing of all possible values of particular test statistic is very computationally intensive already for small N and practically impossible for $N > 20$. Similarly to the bootstrap case, also for the permutation test we need to use approximation — repeating the permutation only B times, where $N < B < N!$. From this procedure we again obtain B estimates of the test statistic distribution function, make a histogram and estimate the critical value for the particular test. More about permutation test can be found, for example, in Good [2005] and specifically about its use in case of change-point problem, for example, in Antoch and Hušková [2001].

1.7 Special distributions

Normal distribution

Let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, then the density of *one-dimensional normal distribution* with parameters μ and σ^2 is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}.$$

If the random variable X follows one-dimensional normal distribution, we denote this $X \sim \mathsf{N}(\mu, \sigma^2)$. The expectation and variance of X are $\mathsf{E}\,X = \mu$ and $\operatorname{var} X = \sigma^2$.

A special case of one-dimensional normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$ is called *standard normal distribution* and we denote Φ its distribution function.

Let $\boldsymbol{X}$ be n-dimensional random vector, $\boldsymbol{\mu}$ be n-dimensional vector of numbers and $\boldsymbol{V}$ be a symmetric positive semi-definite matrix $n \times n$. We say that vector $\boldsymbol{X}$ follows *n-dimensional normal distribution* if

$$\boldsymbol{c}^T\boldsymbol{X} \sim \mathsf{N}\left(\boldsymbol{c}^T\boldsymbol{\mu}, \boldsymbol{c}^T\boldsymbol{V}\boldsymbol{c}\right)$$

for any vector $\boldsymbol{c} \in \mathbb{R}^n$. If the matrix $\boldsymbol{V}$ is nonsingular, the density of n-dimensional normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{V}$ is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{V}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \quad \boldsymbol{x} \in \mathbb{R}^n.$$

The expectation and variance matrix of $\boldsymbol{X}$ are $\mathsf{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\operatorname{var} \boldsymbol{X} = \boldsymbol{V}$.

t-distribution

For $n \geq 1$ the *t-distribution* (or *Student distribution*) with n degrees of freedom is defined by density

$$f(x) = \frac{\Gamma\left(\frac{N+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}},$$

where $\Gamma(.)$ is gamma function defined by $\Gamma(a) = \int_0^\infty x^{a-1}\mathrm{e}^{-x}\,\mathrm{d}x$ for $a > 0$. If the random variable X follows t-distribution with n degrees of freedom, we denote this $X \sim \mathrm{t}_n$. The expectation and variance of X are $\mathsf{E}\,X = 0$ (only if $n > 1$, otherwise it does not exist) and $\operatorname{var} X = \frac{n}{n-2}$ (only if $n > 2$, otherwise it does not exist).

χ^2-distribution

For $n \geq 1$ the *χ^2-distribution* with n degrees of freedom is defined by density

$$f(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}, \quad x > 0.$$

If the random variable X follows χ^2-distribution with n degrees of freedom, we denote this $X \sim \chi^2_n$. The expectation and variance of X are $\mathsf{E}\,X = n$ and $\operatorname{var} X = 2n$.

Bernoulli distribution

Let $p \in (0, 1)$ and consider the set of possible values to be $\{0, 1\}$. The distribution with probabilities $\mathsf{P}(X = 1) = p$ and $\mathsf{P}(X = 0) = 1 - p$ is called *Bernoulli* *distribution* and its expectation and variance are $\mathsf{E}\,X = p$ and var $X = p\,(1-p)$.

Binomial distribution

Let $n \in \mathbb{N}$ and $p \in (0,1)$ and consider the set of possible values to be $\{0, 1, \ldots, n\}$. The distribution with probabilities

$$\mathsf{P}(X = k) = \binom{n}{k} p^k \,(1-p)^{n-k}, \quad \text{for } k = 0, 1, \ldots, n,$$

is called *Binomial distribution.* If the random variable X follows binomial distribution, we denote this $X \sim \mathsf{Bi}(n, p)$. The expectation and variance of X are $\mathsf{E}\,X = np$ and var $X = np(1-p)$.

Multinomial distribution

Let $k \geq 2$, $n \in \mathbb{N}$ and $p_i \in (0,1)$ for $i = 1, \ldots, k$ such that $\sum_{i=1}^{k} p_k = 1$. Denote $\boldsymbol{p} = (p_1, \ldots, p_k)^T$. The distribution of a random vector $\boldsymbol{X} = (X_1, \ldots, X_k)^T$ with probabilities

$$\mathsf{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

where $x_i = 0, 1, \ldots, n$ for $i = 1, \ldots, k$ and $\sum_{i=1}^{k} x_i = n$, is called *multinomial distribution* with parameters n and $\boldsymbol{p}$. The expectation of $\boldsymbol{X}$ is $\mathsf{E}\boldsymbol{X} = (np_1, \ldots, np_k)^T$ and the variance matrix has elements var $X_i = np_i\,(1-p_i)$ and cov $(X_i, X_j) = -np_i p_j$ for $i, j = 1, \ldots, k$ and $i \neq j$.

Chapter 2

Broken-line case

Studying detection of change in broken-line model is often motivated by meteorological or hydrological data. The data about rainfall series in Sahel has served as a motivation for a test of detection of a gradual change in simple linear regression in Jarušková [1998].

First, denote $(X_1, Y_1), \ldots, (X_N, Y_N)$ a random sample from some two-dimensional distribution F ordered by X_i $(X_1 \leq X_2 \leq \cdots \leq X_N)$. A general broken-line linear model with one change point can be defined as following

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 (X_i - X_m) I(i > m) + \varepsilon_i, \quad \text{for } i = 1, \ldots, N,$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^T$ are unknown regression parameters, $\gamma_2 \neq 0$, m is the change point and ε_i, $i = 1, \ldots, N$ are independent identically distributed random variables with zero mean and finite variance σ^2. The vector $\boldsymbol{X} = (X_1, \ldots, X_N)^T$ can be either vector of random variables (as introduced in above model) or vector of constants. In this thesis we deal only with the case of constants in specific form $X_i = \frac{i}{N}$.

The situation of independent identically distributed random variables against simple broken line model with zero slope before an unknown point of change and non-zero slope after the point of change with special case of $X_i = i$ and ε_i such that $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$ and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1, \ldots, N$ was considered in Jarušková [1998]. The null and alternative hypotheses for this case can be written as

$$H_0 : Y_i = \alpha_0 + \varepsilon_i, \quad \text{for } i = 1, \ldots, N$$

vs.

$$\begin{aligned} H_1 : \quad & Y_i = \alpha_0 + \varepsilon_i, \quad \text{for } i = 1, \ldots, m \\ & Y_i = \alpha_0 + \alpha_2 (i - m) + \varepsilon_i, \quad \text{for } i = m+1, \ldots, N \end{aligned}$$

for some $m \in \{1, \ldots, N-1\}$, where α_0 and α_2 are regression coefficients, $\alpha_2 \neq 0$, and ε_i, $i = 1, \ldots, N$ are independent identically distributed random variables such that $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$ and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1, \ldots, N$.

This chapter deals with little broader model which considers linear trend before the change point instead of constant mean while keeping the condition of continuity in the change point and $X_i = \frac{i}{N}$ and ε_i such that $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$

and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1, \ldots, N$. The hypothesis for this case can be written as

$$
\begin{aligned}
& H_0 : Y_i = \beta_0 + \beta_1 \frac{i}{N} + \varepsilon_i, \quad \text{for } i = 1, \ldots, N \\
& \text{vs.} \\
H_1 : \quad & Y_i = \beta_0 + \beta_1 \frac{i}{N} + \varepsilon_i, \quad \text{for } i = 1, \ldots, m \\
& Y_i = \beta_0 + \beta_1 \frac{i}{N} + \beta_2 \frac{i-m}{N} + \varepsilon_i, \quad \text{for } i = m+1, \ldots, N
\end{aligned}
\tag{2.1}
$$

for some $m \in \{2, \ldots, N-2\}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ is vector of regression coefficients, $\beta_2 \neq 0$, and ε_i, $i = 1, \ldots, N$ are independent identically distributed random variables such that $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$ and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1, \ldots, N$.

2.1 Known change point

In case of known change point m the above hypothesis can be rephrased as

$$
\begin{gathered}
H_0 : \beta_2 = 0 \\
\text{vs.} \\
H_1 : \beta_2 \neq 0
\end{gathered}
\tag{2.2}
$$

in the model

$$
Y_i = \beta_0 + \beta_1 \frac{i}{N} + \beta_2 \frac{i-m}{N} I(i > m) + \varepsilon_i, \quad \text{for } i = 1, \ldots, N, \tag{2.3}
$$

where m is the change point and ε_i, $i = 1, \ldots, N$ are independent identically distributed random variables such that $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$ and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1, \ldots, N$.

Test statistic

Natural test statistic for this case is a normalized estimator of parameter β_2. Let's denote $\hat{\beta}_2^{(m)}$ the least squares estimator of β_2 from the model (2.3).

This estimator can be expressed as

$$
\hat{\beta}_2^{(m)} = \frac{\sum_{i=m}^{N} \frac{i-m}{N} \left(Y_i - \bar{Y}_N\right) - \frac{\sum_{i=m}^{N} \frac{i-m}{N} \left(\frac{i}{N} - \frac{N+1}{2N}\right)}{\sum_{i=1}^{N} \left(\frac{i}{N} - \frac{N+1}{2N}\right)^2} \sum_{i=1}^{N} \left(\frac{i}{N} - \frac{N+1}{2N}\right) \left(Y_i - \bar{Y}_N\right)}{\sum_{i=m}^{N} \left(\frac{i-m}{N}\right)^2 - \frac{1}{N} \left(\sum_{i=m}^{N} \frac{i-m}{N}\right)^2 - \frac{\left(\sum_{i=m}^{N} \frac{i-m}{N} \left(\frac{i}{N} - \frac{N+1}{2N}\right)\right)^2}{\sum_{i=1}^{N} \left(\frac{i}{N} - \frac{N+1}{2N}\right)^2}} \tag{2.4}
$$

and its variance is

$$
\operatorname{var} \hat{\beta}_2^{(m)} = \sigma^2 \left[\sum_{i=m}^{N} \left(\frac{i-m}{N}\right)^2 - \frac{1}{N} \left(\sum_{i=m}^{N} \frac{i-m}{N}\right)^2 - \frac{\left(\sum_{i=m}^{N} \frac{i-m}{N} \left(\frac{i}{N} - \frac{N+1}{2N}\right)\right)^2}{\sum_{i=1}^{N} \left(\frac{i}{N} - \frac{N+1}{2N}\right)^2} \right] \tag{2.5}
$$

Let's denote

$$R_I^{(m)} = \frac{\hat{\beta}_2^{(m)}}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}} \tag{2.6}$$
$$= \frac{\sum_{i=m}^{N} \frac{i-m}{N}\left(Y_i - \bar{Y}_N\right) - \frac{\sum_{i=m}^{N} \frac{i-m}{N}\left(\frac{i}{N} - \frac{N+1}{2N}\right)}{\sum_{i=1}^{N}\left(\frac{i}{N} - \frac{N+1}{2N}\right)^2} \sum_{i=1}^{N}\left(\frac{i}{N} - \frac{N+1}{2N}\right)\left(Y_i - \bar{Y}_N\right)}{\sigma\sqrt{\sum_{i=m}^{N}\left(\frac{i-m}{N}\right)^2 - \frac{1}{N}\left(\sum_{i=m}^{N}\frac{i-m}{N}\right)^2 - \frac{\left(\sum_{i=m}^{N}\frac{i-m}{N}\left(\frac{i}{N} - \frac{N+1}{2N}\right)\right)^2}{\sum_{i=1}^{N}\left(\frac{i}{N} - \frac{N+1}{2N}\right)^2}}}$$

the test statistic of our interest for testing the hypothesis (2.2).

Distribution of test statistic under $\mathbf{H}_0$: Normally distributed random errors with known $\boldsymbol{\sigma}^2$

In simple case of normally distributed random errors with known σ^2 the distribution of test statistic (2.6) is known:

$$R_I^{(m)} \sim \mathsf{N}(0,1) \tag{2.7}$$

[see for example Anděl, 2005, pages 81–83] and we reject the null hypothesis (2.2) on the level of significance α if

$$\left|R_I^{(m)}\right| > z_{1-\alpha/2}, \tag{2.8}$$

where $z_{1-\alpha/2}$ is $100(1-\alpha/2)\%$ quantile of standard normal distribution.

Distribution of test statistic under $\mathbf{H}_0$: Normally distributed random errors with unknown $\boldsymbol{\sigma}^2$

If we keep normality of the random errors, but assume σ^2 to be unknown, we cannot compute the var $\hat{\beta}_2^{(m)}$ directly as it depends on parameter σ^2. The normalized residual sum of squares

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\left(\hat{\varepsilon}_i\right)^2}{n-3} = \frac{\sum_{i=1}^{N}\left(Y_i - \hat{\beta}_0^{(m)} - \hat{\beta}_1^{(m)}\frac{i}{N} - \hat{\beta}_2^{(m)}\frac{i-m}{N}I(i>m)\right)^2}{n-3}$$

is an unbiased estimator of σ^2. Denote $\widehat{\operatorname{var}}\,\hat{\beta}_2^{(m)}$ the var $\hat{\beta}_2^{(m)}$ with σ^2 replaced with $\hat{\sigma}^2$. Since

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-3}$$

it holds

$$R_{II}^{(m)} = \frac{\hat{\beta}_2^{(m)}}{\widehat{\operatorname{var}}\,\hat{\beta}_2^{(m)}} \sim \mathrm{t}_{N-3} \tag{2.9}$$

[see for example Anděl, 2005, pages 81–83] and we reject the null hypothesis (2.2) on the level of significance α if

$$\left|R_{II}^{(m)}\right| > \mathrm{t}_{N-3}\left(1-\alpha/2\right), \tag{2.10}$$

where $\mathrm{t}_{N-3}(1-\alpha/2)$ is $100(1-\alpha/2)\%$ quantile of t-distribution with $N-3$ degrees of freedom.

Distribution of test statistic under H_0: General independent random errors

Now let's replace the assumption of normality of the random errors ε_i, $i = 1, \ldots, N$ with holding simple conditions $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$ and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1, \ldots, N$ (while ε_i are independent for $i = 1, \ldots, N$). In such case we cannot calculate the exact distribution of $R_I^{(m)}$ and $R_{II}^{(m)}$. However, the above conditions on ε_i, $i = 1, \ldots, N$ are sufficient for assumptions of the Central Limit Theorem (Theorem A.1) to be met, so for known σ^2 it holds

$$R_I^{(m)} \xrightarrow{d} \mathsf{N}(0,1), \quad \text{for } N \to \infty,$$

and we reject the null hypothesis on asymptotical level α if

$$\left| R_I^{(m)} \right| > z_{1-\alpha/2},$$

where $z_{1-\alpha/2}$ is $100(1-\alpha/2)\%$ quantile of standard normal distribution.

Since the estimator $\widehat{\mathrm{var}}\,\hat{\beta}_2^{(m)}$ is a consistent estimator of $\mathrm{var}\,\hat{\beta}_2^{(m)}$ we can replace $\mathrm{var}\,\hat{\beta}_2^{(m)}$ with it while keeping the asymptotic normal distribution (as a consequence of the Crámer-Slutsky Theorem — Theorem A.2), so

$$R_{II}^{(m)} \xrightarrow{d} \mathsf{N}(0,1), \quad \text{for } N \to \infty,$$

and we reject the null hypothesis on asymptotical level α if

$$\left| R_{II}^{(m)} \right| > z_{1-\alpha/2},$$

where $z_{1-\alpha/2}$ is again $100(1-\alpha/2)\%$ quantile of standard normal distribution.

Test power

The test power can be easily computed in case of normally distributed random errors. Denote $\pi_I^{(m)}(\boldsymbol{\beta})$ the power function of test (2.8) with known σ^2 and $\pi_{II}^{(m)}(\boldsymbol{\beta}, \sigma^2)$ the power function of test (2.10) with unknown σ^2, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, then

$$
\begin{aligned}
\pi_I^{(m)}(\boldsymbol{\beta}) &= \mathsf{P}\left(\frac{\left|\hat{\beta}_2^{(m)}\right|}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}} > z_{1-\alpha/2}\right) \\
&= \mathsf{P}\left(\frac{\hat{\beta}_2^{(m)}-\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}} > z_{1-\alpha/2} - \frac{\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}}\right) \\
&\quad + \mathsf{P}\left(\frac{\hat{\beta}_2^{(m)}-\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}} \leq -z_{1-\alpha/2} - \frac{\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}}\right) \\
&= 1-\Phi\left(z_{1-\alpha/2} - \frac{\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}}\right) \\
&\quad + \Phi\left(-z_{1-\alpha/2} - \frac{\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}}\right),
\end{aligned} \tag{2.11}
$$

where $\Phi(.)$ is the distribution function of standard normal distribution $\mathsf{N}(0,1)$, and

$$
\begin{aligned}
\pi_{II}^{(m)}(\boldsymbol{\beta},\sigma^2) &= \mathsf{P}\left(\frac{\left|\hat{\beta}_2^{(m)}\right|}{\sqrt{\widehat{\operatorname{var}}\,\hat{\beta}_2^{(m)}}} > \mathrm{t}_{N-3}\left(1-\alpha/2\right)\right) \\
&= \mathsf{P}\left(\frac{\hat{\beta}_2^{(m)}-\beta_2}{\sqrt{\widehat{\operatorname{var}}\,\hat{\beta}_2^{(m)}}} > \mathrm{t}_{N-3}\left(1-\alpha/2\right) - \frac{\beta_2}{\sqrt{\widehat{\operatorname{var}}\,\hat{\beta}_2^{(m)}}}\right) \\
&\quad + \mathsf{P}\left(\frac{\hat{\beta}_2^{(m)}-\beta_2}{\sqrt{\widehat{\operatorname{var}}\,\hat{\beta}_2^{(m)}}} \leq -\mathrm{t}_{N-3}\left(1-\alpha/2\right) - \frac{\beta_2}{\sqrt{\widehat{\operatorname{var}}\,\hat{\beta}_2^{(m)}}}\right) \\
&\approx 1-\Phi\left(\mathrm{t}_{N-3}\left(1-\alpha/2\right) - \frac{\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}}\right) \\
&\quad + \Phi\left(-\mathrm{t}_{N-3}\left(1-\alpha/2\right) - \frac{\beta_2}{\sqrt{\operatorname{var}\,\hat{\beta}_2^{(m)}}}\right),
\end{aligned} \tag{2.12}
$$

according to the Central Limit Theorem (Theorem A.1) for large N.

In case of omitting the condition of normality and introducing the conditions $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$ and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1,\ldots,N$ (while ε_i are independent), the power function needs to be approximated. Denote $\pi_I^{(m)*}(\boldsymbol{\beta})$ the power function for known σ^2 case and $\pi_{II}^{(m)*}(\boldsymbol{\beta},\sigma^2)$ the power function for unknown σ^2 case. In case of known σ^2, the power is equal to (2.11), but we do not know the exact distribution this time. We can approximate it using the Central Limit Theorem

(Theorem A.1):

$$\pi_I^{(m)*}(\boldsymbol{\beta}) \approx 1 - \Phi\left(z_{1-\alpha/2} - \frac{\beta_2}{\sqrt{\text{var } \hat{\beta}_2^{(m)}}}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{\beta_2}{\sqrt{\text{var } \hat{\beta}_2^{(m)}}}\right),$$

where $\Phi(.)$ is the distribution function of standard normal distribution $\mathsf{N}(0,1)$. Similarly for the case of unknown σ^2, the power is equal to (2.12), but we do not know the exact distribution. The approximation by using the Central Limit Theorem (Theorem A.1) can be used in this case as well:

$$\pi_{II}^{(m)*}(\boldsymbol{\beta}, \sigma^2) \approx 1 - \Phi\left(\mathrm{t}_{N-3}(1-\alpha/2) - \frac{\beta_2}{\sqrt{\text{var } \hat{\beta}_2^{(m)}}}\right) + \Phi\left(-\mathrm{t}_{N-3}(1-\alpha/2) - \frac{\beta_2}{\sqrt{\text{var } \hat{\beta}_2^{(m)}}}\right),$$

where $\Phi(.)$ is the distribution function of standard normal distribution $\mathsf{N}(0,1)$.

2.2 Unknown change point

The previous Section dealt with situation of known change point which led us to testing a parameter in simple linear regressions to be zero. However, the change point is often unknown and in such case we cannot use the approach described in previous section. In this Section we look at how the situation changes in case of unknown change point.

Test statistic

The alternative hypothesis (2.1) says that there exists a change point $m < N$ for which the parameter β_2 is non-zero. If the change point is unknown, one of our options is to search for a maximum of the test statistic (2.7) over possible values of m. Hence the test statistic of our interest is the maximum-type statistic in this form:

$$R_{III} = \max_{2 \le j \le N-2} \left\{ \frac{\left|\hat{\beta}_2^{(j)}\right|}{\sqrt{\widehat{\text{var}} \ \hat{\beta}_2^{(j)}}} \right\} \tag{2.13}$$

where expressions for $\hat{\beta}_2^{(j)}$ and $\sqrt{\widehat{\text{var}} \ \hat{\beta}_2^{(j)}}$ can be found in (2.4) and (2.5).

Distribution of test statistic under $\mathbf{H}_0$

Let's look at distribution of this test statistic under the null hypothesis. One of the possible approximations of the distribution for general case of random errors ε_i with $\mathsf{E}\,\varepsilon_i = 0$, $\mathsf{E}\,\varepsilon_i^2 = \sigma^2$ and $\mathsf{E}\,|\varepsilon_i|^{2+\delta} < \infty$ for $i = 1, \ldots, N$ can be found in Antoch et al. [2000, page 33]:

$$\mathsf{P}\left(\max_{2 \le j \le N-2} \left\{ \frac{\left|\hat{\beta}_2^{(j)}\right|}{\sqrt{\widehat{\operatorname{var}}\, \hat{\beta}_2^{(j)}}} \right\} > \frac{x + b_{N,3}}{a_N} \right) \approx 1 - \exp\left\{-4\,\mathrm{e}^{-x}\right\},$$

for $x \in \mathbb{R}$ and large N, where

$$a_N = \sqrt{2\,\log\,\log\,N}$$

and

$$b_{N,3} = 2\,\log\,\log\,N + \log\,\frac{\sqrt{3}}{4\pi}.$$

Denote

$$c_N = \frac{x + b_{N,3}}{a_N}$$

the critical value of the test. This approximation is obtained using extreme value theory, which usually leads to approximations that converge to the exact distribution very slowly [see Csörgő and Horváth, 1997].

Test power

The approximation of power for this case would be very complex and so it is not studied in this thesis.

2.3 Bootstrap and permutation test

The bootstrap algorithm and the procedure of permutation test for the broken line case are described in this Section. The algorithms for both bootstrap and permutations and for all three test statistics that were introduced in Sections 2.1 and 2.2 are similar, hence we describe only one of them and then explain the differences for the other cases. Semi-parametric versions of bootstrap and permutation test are used, it means that we are not sampling directly the values of Y_i, $i = 1, \ldots, N$, but the random errors from model under the null hypothesis.

The bootstrap algorithm for test statistic (2.7):

1. fit the model under the null hypothesis (it means $Y_i = \beta_0 + \beta_1\,X_i + \varepsilon_i$, $i = 1, \ldots, N$) with the original data $(X_1, Y_1), \ldots, (X_N, Y_N)$ and estimate $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\varepsilon}_i$, $i = 1, \ldots, N$ (where $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1\,X_i$)
2. for $b = 1, \ldots, B$, where B is the chosen number of bootstrap repetitions
   (a) sample N values from $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N$ with replacement and denote $\hat{\varepsilon}^*_{1,b}, \ldots, \hat{\varepsilon}^*_{N,b}$

(b) compute values of $Y^*_{1,b}, \dots, Y^*_{N,b}$ where $Y^*_{i,b} = \hat\beta_0 + \hat\beta_1 X_i + \hat\varepsilon^*_{i,b}$, $i = 1, \dots, N$

(c) fit the model (2.3) with bootstrapped values $(X_1, Y^*_{1,b}), \dots, (X_N, Y^*_{N,b})$ and estimate $\hat\beta_2^{(m)}$ (denote $\hat\beta_{2,b}^{(m)*}$)

(d) compute the values of $\sqrt{\operatorname{var} \hat\beta_{2,b}^{(m)*}}$ and the test statistic $R_I^{(m)}$ for this sample and denote it $R_{I,b}^{(m)*}$

3. find the $\lfloor(1-\alpha)B\rfloor$-th smallest value from $R_{I,1}^{(m)*}, \dots, R_{I,B}^{(m)*}$ and denote t_{boot}
4. the t_{boot} is the critical value for the test.

In case of permutation test the only difference is in step 2a:

2a* sample N values from $\hat\varepsilon_1, \dots, \hat\varepsilon_N$ *without* replacement and denote $\hat\varepsilon^*_{1,b}, \dots, \hat\varepsilon^*_{N,b}$

Both the algorithms for bootstrap and permutation test are similar for the test statistics (2.9) and (2.13).

The algorithm for simulating the power of these tests is again similar for all three test statistics and both permutation and bootstrap tests, so only one version is described in detail.

The simulated power for bootstrap test with test statistic $R_I^{(m)}$ can be computed using this algorithm:

i. generate R data sets under the alternative hypothesis with particular values of β_0, β_1, β_2, N and m

ii. for each of these data sets compute the value of the test statistic $R_I^{(m)}$ and denote $R_I^{(m),r}$ for $r = 1, \dots, R$

iii. the simulated power is equal to portion of $R_I^{(m),r}$ for which $|R_I^{(m),r}| > t_{boot}$ where t_{boot} is critical value for the particular test.

Chapter 3

Discrete variable case

The change-point problem for case of discrete variables is studied on example from ecology in López et al. [2010]. The authors of this article consider one plant species' presence along a line transect. The distribution along this line transect is inhomogeneous and there can be multiple change points that cut the line transect into multiple homogeneous parts. For simplicity the article deals only with case of one such a change point. However, this article deals only with estimation of the parameters, not with testing if there is really some change. In this chapter we deal with testing of change for two situations. The simpler one looks at a Bernoulli distributed variables (Section 3.1), the more complex one is the extension for discrete variables with more than two possible values (Section 3.2).

3.1 Bernoulli distribution

Let $X_1, \ldots, X_N$ be independent Bernoulli distributed random variables. Let's consider $X_1, \ldots, X_K$ to be a random sample from Bernoulli distribution with parameter p and $X_{K+1}, \ldots, X_N$ to be a random sample from Bernoulli distribution with parameter q where $p,\ q \in (0,1)$. The hypothesis about the change in distribution at point K can be formulated as

$$\begin{gathered} H_0 : p = q \\ \text{vs.} \\ H_1 : p \neq q. \end{gathered} \tag{3.1}$$

This change point is usually unknown. The true point of change K depends on N so we should properly write K_N. However, since there is no confusion, we will omit the index N and will write only K. For avoiding the extreme situations of $K \to 0$ and $N - K \to 0$ for $N \to \infty$ we assume in the whole section that there exists some $\kappa \in (0,1)$ such that $\frac{K}{N} \to \kappa$ for $N \to \infty$. In Section 3.1.1 we consider the change point K to be known and describe the testing process in this simple case, in Section 3.1.2 the extension for unknown change point K is described.

3.1.1 Known change point

Test statistic

The above situation with known change point K is equivalent to two-sample problem. The common approach could be using the maximum likelihood ratio statistic (see for example Chen and Gupta [2000] for particular cases of Binomial and Poisson distribution). However, this method would lead to extreme value distribution of the test statistic in case of extension to the unknown change point, hence we have chosen another approach to this problem — we will study the standardized difference of the maximum likelihood estimates for p and q. Under the null hypothesis (3.1) this difference is small and goes to 0 with increasing number of observations, while under the alternative hypothesis this would approach the $+\infty$ or $-\infty$.

The maximum likelihood estimators for parameters p and q are in such case

$$\hat{p}^{(K)} = \frac{1}{K}\sum_{i=1}^{K} X_i \tag{3.2}$$

and

$$\hat{q}^{(K)} = \frac{1}{N-K}\sum_{i=K+1}^{N} X_i. \tag{3.3}$$

These estimators are unbiased and their expectations and variances are

$$\begin{aligned}
\mathsf{E}\hat{p}^{(K)} &= \mathsf{E}\left(\frac{1}{K}\sum_{i=1}^{K} X_i\right) = \frac{1}{K}\sum_{i=1}^{K}\mathsf{E}X_i = \mathsf{E}X_1 = p,\\
\mathsf{E}\hat{q}^{(K)} &= \mathsf{E}\left(\frac{1}{N-K}\sum_{i=K+1}^{N} X_i\right) = \frac{1}{N-K}\sum_{i=K+1}^{N}\mathsf{E}X_i = \mathsf{E}X_N = q
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{var}\hat{p}^{(K)} &= \operatorname{var}\left(\frac{1}{K}\sum_{i=1}^{K} X_i\right) = \frac{1}{K^2}\sum_{i=1}^{K}\operatorname{var}X_i = \\
&= \frac{1}{K}\operatorname{var}X_1 = \frac{1}{K}p(1-p),\\
\operatorname{var}\hat{q}^{(K)} &= \operatorname{var}\left(\frac{1}{N-K}\sum_{i=K+1}^{N} X_i\right) = \frac{1}{(N-K)^2}\sum_{i=K+1}^{N}\operatorname{var}X_i = \\
&= \frac{1}{N-K}\operatorname{var}X_N = \frac{1}{N-K}q(1-q).
\end{aligned}$$

Denote $Z^{(K)} = \hat{p}^{(K)} - \hat{q}^{(K)}$. As all X_i are independent, $\hat{p}^{(K)}$ and $\hat{q}^{(K)}$ are independent, too, and the expectation and variance of variable $Z^{(K)}$ are

$$\mathsf{E}Z^{(K)} = \mathsf{E}\hat{p}^{(K)} - \mathsf{E}\hat{q}^{(K)} = p - q$$

and

$$\begin{aligned}\operatorname{var} Z^{(K)} &= \operatorname{var} \hat{p}^{(K)} + \operatorname{var} \hat{q}^{(K)} = \\ &= \frac{1}{K} p(1-p) + \frac{1}{N-K} q(1-q).\end{aligned}$$

Under the null hypothesis (3.1) the expectation of $Z^{(K)}$ is equal to 0 and the variance is $S^{(K)} = \frac{N}{K(N-K)} p(1-p)$. Since X_i, $i = 1, \dots, N$ are Bernoulli distributed, the assumptions of the Central Limit Theorem (Theorem A.1) are met and we obtain

$$T^{(K)} = \frac{Z^{(K)}}{\sqrt{S^{(K)}}} \xrightarrow{d} \mathsf{N}(0,1), \quad \text{for } N \to \infty,\ K \to \infty,\ \frac{K}{N} \to \kappa. \tag{3.4}$$

Distribution of test statistic under $\mathbf{H}_0$

The variance of the variable $Z^{(K)}$ depends on unknown p, so we need to estimate it. We use two different possible estimators for $S^{(K)}$ under the null hypothesis. We can estimate the variance using the estimator

$$\tilde{S}^{(K)} = \frac{N}{K(N-K)} \tilde{p}_N (1 - \tilde{p}_N), \tag{3.5}$$

where

$$\tilde{p}_N = \frac{1}{N} \sum_{i=1}^{N} X_i, \tag{3.6}$$

or using the estimator

$$\hat{S}^{(K)} = \frac{1}{K} \hat{p}^{(K)} (1 - \hat{p}^{(K)}) + \frac{1}{N-K} \hat{q}^{(K)} (1 - \hat{q}^{(K)}), \tag{3.7}$$

where $\hat{p}^{(K)}$ and $\hat{q}^{(K)}$ are the estimators defined in (3.2) and (3.3). Both (3.5) and (3.7) are consistent estimators of $S^{(K)}$, so according to the Cramér-Slutsky Theorem (Theorem A.2) we can replace the $S^{(K)}$ in (3.4) with either (3.5) or (3.7) while keeping the asymptotical normality of the test statistic:

$$T_I^{(K)} = \frac{Z^{(K)}}{\sqrt{\tilde{S}^{(K)}}} \xrightarrow{d} \mathsf{N}(0,1), \quad \text{for } N \to \infty,\ K \to \infty,\ \frac{K}{N} \to \kappa \tag{3.8}$$

and

$$T_{II}^{(K)} = \frac{Z^{(K)}}{\sqrt{\hat{S}^{(K)}}} \xrightarrow{d} \mathsf{N}(0,1), \quad \text{for } N \to \infty,\ K \to \infty,\ \frac{K}{N} \to \kappa. \tag{3.9}$$

Since the value of the test statistic (3.8) is large under the alternative hypothesis (see next paragraph), we reject the null hypothesis (3.1) on the asymptotical significance level α if

$$|T_I^{(K)}| > z_{1-\alpha/2}, \tag{3.10}$$

where $z_{1-\alpha/2}$ is $100(1-\alpha/2)\%$ quantile of the standard normal distribution.

Similarly for test using test statistic (3.9) we reject the null hypothesis on the asymptotical significance level α if

$$|T_{II}^{(K)}| > z_{1-\alpha/2}. \tag{3.11}$$

Distribution of test statistic under $\mathbf{H}_1$

Now let's look at the distribution of the test statistics (3.8) and (3.9) under the alternative hypothesis. For fixed alternative $p \neq q$ the expectation of $Z^{(K)}$ is equal to difference between p and q which is non-zero number between -1 and 1. For the parameter estimators (3.2), (3.3) and (3.6) under the alternative hypothesis it holds

$$\mathsf{E}_{H_1}\hat{p}^{(K)} = p,$$

$$\mathsf{E}_{H_1}\hat{q}^{(K)} = q$$

and

$$\tilde{p}_N = \frac{1}{N}\left(\sum_{i=1}^{K} X_i + \sum_{i=K+1}^{N} X_i\right) \xrightarrow{\mathsf{P}} \kappa\, p + (1-\kappa)\, q, \quad \text{for } N \to \infty.$$

With respect to our assumptions, all these three numbers are non-zero and bounded. For (3.5) and (3.7) it holds

$$N\,\tilde{S}^{(K)} \xrightarrow{\mathsf{P}} \frac{1}{1-\kappa}\,p + \frac{1}{\kappa}\,q - \frac{\left[\kappa\, p + (1-\kappa)\, q\right]^2}{\kappa\,(1-\kappa)} \quad \text{for } N \to \infty \tag{3.12}$$

and

$$N\,\hat{S}^{(K)} \xrightarrow{\mathsf{P}} \frac{1}{\kappa}\,p\,(1-p) + \frac{1}{1-\kappa}\,q\,(1-q) \quad \text{for } N \to \infty. \tag{3.13}$$

The expression on the right-hand side of (3.12) is a sum of three multiples of non-zero bounded numbers which is again non-zero bounded number and

$$\left|T_I^{(K)}\right| = \left|\frac{\sqrt{N}\,Z^{(K)}}{\sqrt{N\,\tilde{S}^{(K)}}}\right| \xrightarrow{\mathsf{P}} +\infty, \quad \text{for } N \to \infty$$

under the alternative hypothesis.

Similarly, the expression on right-hand side of (3.13) is a sum of two multiples of non-zero bounded numbers which is again non-zero bounded number and

$$\left|T_{II}^{(K)}\right| = \left|\frac{\sqrt{N}\,Z^{(K)}}{\sqrt{N\,\hat{S}^{(K)}}}\right| \xrightarrow{\mathsf{P}} +\infty, \quad \text{for } N \to \infty$$

under the alternative hypothesis.

Test power

The approximation of the power for these two tests can be easily computed. Denote $\xi_I^{(K)}(p,q)$ the power function for the test (3.10), then

$$\begin{aligned}
\xi_I^{(K)}(p,q) &= \mathsf{P}\left(\frac{\sqrt{N}\left|\hat{p}^{(K)}-\hat{q}^{(K)}\right|}{\sqrt{\frac{1}{\kappa\,(1-\kappa)}\tilde{p}_N\,(1-\tilde{p}_N)}} > z_{1-\alpha/2}\right) \\
&= \mathsf{P}\left(\frac{\sqrt{N}\left(\hat{p}^{(K)}-\hat{q}^{(K)}-p+q\right)}{\sqrt{\frac{1}{\kappa}p\,(1-p)+\frac{1}{1-\kappa}q\,(1-q)}}\right. \\
&\qquad\qquad \left. > \frac{\sqrt{\frac{1}{\kappa\,(1-\kappa)}\tilde{p}_N\,(1-\tilde{p}_N)}\,z_{1-\alpha/2}-\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\,(1-p)+\frac{1}{1-\kappa}q\,(1-q)}}\right) \\
&+ \mathsf{P}\left(\frac{\sqrt{N}\left(\hat{p}^{(K)}-\hat{q}^{(K)}-p+q\right)}{\sqrt{\frac{1}{\kappa}p\,(1-p)+\frac{1}{1-\kappa}q\,(1-q)}}\right. \\
&\qquad\qquad \left. \le \frac{-\sqrt{\frac{1}{\kappa\,(1-\kappa)}\tilde{p}_N\,(1-\tilde{p}_N)}\,z_{1-\alpha/2}-\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\,(1-p)+\frac{1}{1-\kappa}q\,(1-q)}}\right).
\end{aligned}$$

Using the Central Limit Theorem (Theorem A.1) we can approximate the power by

$$\begin{aligned}
\xi_I^{(K)}(p,q) &\approx 1-\Phi\left(\frac{\sqrt{\frac{(\kappa p+(1-\kappa)q)\,(1-\kappa p-(1-\kappa)q)}{\kappa\,(1-\kappa)}}\,z_{1-\alpha/2}-\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\,(1-p)+\frac{1}{1-\kappa}q\,(1-q)}}\right) \\
&+\Phi\left(\frac{-\sqrt{\frac{(\kappa p+(1-\kappa)q)\,(1-\kappa p-(1-\kappa)q)}{\kappa\,(1-\kappa)}}\,z_{1-\alpha/2}-\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\,(1-p)+\frac{1}{1-\kappa}q\,(1-q)}}\right).
\end{aligned}$$

Looking closer at the last two elements in the above approximation we can see, that first part of numerator and whole denominator are bounded numbers. The remaining part is approaching either $+\infty$ (for $p<q$) or $-\infty$ (for $p>q$) for $N\to\infty$ and so both elements are going to 0, or both elements are going to 1 and whole expression for approximated power is approaching 1 for $N\to\infty$.

Similarly, denote $\xi_{II}^{(K)}(p,q)$ the power function for test (3.11), then

$$\begin{aligned}
\xi_{II}^{(K)}(p,q) &= \mathsf{P}\left(\frac{\sqrt{N}\left|\hat{p}^{(K)}-\hat{q}^{(K)}\right|}{\sqrt{\frac{1}{\kappa}\hat{p}^{(K)}\left(1-\hat{p}^{(K)}\right)+\frac{1}{1-\kappa}\hat{q}^{(K)}\left(1-\hat{q}^{(K)}\right)}} > z_{1-\alpha/2}\right) \\
&= \mathsf{P}\left(\frac{\sqrt{N}\left(\hat{p}^{(K)}-\hat{q}^{(K)}-p+q\right)}{\sqrt{\frac{1}{\kappa}p\left(1-p\right)+\frac{1}{1-\kappa}q\left(1-q\right)}}\right. \\
&\qquad \left. > \frac{\sqrt{\frac{1}{\kappa}\hat{p}^{(K)}\left(1-\hat{p}^{(K)}\right)+\frac{1}{1-\kappa}\hat{q}^{(K)}\left(1-\hat{q}^{(K)}\right)}\, z_{1-\alpha/2}-\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\left(1-p\right)+\frac{1}{1-\kappa}q\left(1-q\right)}}\right) \\
&+ \mathsf{P}\left(\frac{\sqrt{N}\left(\hat{p}^{(K)}-\hat{q}^{(K)}-p+q\right)}{\sqrt{\frac{1}{\kappa}p\left(1-p\right)+\frac{1}{1-\kappa}q\left(1-q\right)}}\right. \\
&\qquad \left. \leq \frac{-\sqrt{\frac{1}{\kappa}\hat{p}^{(K)}\left(1-\hat{p}^{(K)}\right)+\frac{1}{1-\kappa}\hat{q}^{(K)}\left(1-\hat{q}^{(K)}\right)}\, z_{1-\alpha/2}-\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\left(1-p\right)+\frac{1}{1-\kappa}q\left(1-q\right)}}\right).
\end{aligned}$$

Again, using the Central Limit Theorem (Theorem A.1) we can approximate the power by

$$\begin{aligned}
\xi_{II}^{(K)}(p,q) \approx 1 - \Phi &\left(z_{1-\alpha/2}-\frac{\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\left(1-p\right)+\frac{1}{1-\kappa}q\left(1-q\right)}}\right) \\
+ \quad \Phi &\left(-z_{1-\alpha/2}-\frac{\sqrt{N}(p-q)}{\sqrt{\frac{1}{\kappa}p\left(1-p\right)+\frac{1}{1-\kappa}q\left(1-q\right)}}\right).
\end{aligned}$$

Looking at the last two elements in the above approximation we can see that the first part is a bounded number, and that the remaining part is approaching either $+\infty$ or $-\infty$ for $N \to \infty$. Hence also in this case both elements are going to 0, or both elements are going to 1 and the approximated power is approaching 1 for $N \to \infty$.

3.1.2 Unknown change point

Test statistic

In this section we introduce extension of the test statistic (3.4) from previous section for unknown change point. First denote

$$D_i = \frac{X_i - p}{\sqrt{p\left(1-p\right)}},$$

for $i = 1, \ldots, N$. Under the null hypothesis (3.1) the random variables X_i, $i = 1, \ldots, N$ are independent identically distributed with Bernoulli distribution with expectation p and variance $p\left(1-p\right)$ and so the random variables D_i, $i = 1, \ldots, N$ are also independent and has expectation 0 and variance 1.

We define the process $\{D_N(t),\ t \in (0,1)\}$ by

$$D_N(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nt \rfloor} D_i + \frac{Nt - \lfloor Nt \rfloor}{\sqrt{N}} D_{\lfloor Nt \rfloor + 1}.$$

Then according to Donsker's theorem (Theorem A.3) we have

$$\{D_N(t),\ t \in (0,1)\} \xrightarrow{d} \{W(t),\ t \in (0,1)\}, \quad \text{for } N \to \infty,$$

and

$$D_N(1) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_i \xrightarrow{d} \mathsf{N}(0,1) = W(1), \quad \text{for } N \to \infty,$$

where $\{W(t), t \in (0,1)\}$ is a Wiener process.

Denote $R_N(t) = D_N(t) - t \cdot D_N(1)$ for $t \in (0,1)$, then according to Corollary A.5

$$\{R_N(t),\ t \in (0,1)\} \xrightarrow{d} \{B(t),\ t \in (0,1)\}, \quad \text{for } N \to \infty$$

where $\{B(t),\ t \in (0,1)\}$ is a Brownian bridge.

We can rewrite the process $\{R_N(t),\ t \in (0,1)\}$:

$$\begin{aligned}
R_N(t) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nt \rfloor} \frac{X_i - p}{\sqrt{p\,(1-p)}} + \frac{Nt - \lfloor Nt \rfloor}{\sqrt{N}} \frac{X_{\lfloor Nt \rfloor + 1} - p}{\sqrt{p\,(1-p)}} \\
&\quad - \frac{t}{\sqrt{N}} \sum_{i=1}^{N} \frac{X_i - p}{\sqrt{p\,(1-p)}} \pm \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nt \rfloor} \frac{\bar{X}_N - p}{\sqrt{p\,(1-p)}} \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nt \rfloor} \frac{X_i - \bar{X}_N}{\sqrt{p,(1-p)}} + \frac{Nt - \lfloor Nt \rfloor}{\sqrt{N}\sqrt{p\,(1-p)}} \left(X_{\lfloor Nt \rfloor + 1} - p\right) \\
&\quad - \frac{Nt}{\sqrt{N}\sqrt{p\,(1-p)}} \bar{X}_N + \frac{Nt}{\sqrt{N}\sqrt{p\,(1-p)}} p + \frac{\lfloor Nt \rfloor}{\sqrt{N}\sqrt{p\,(1-p)}} \left(\bar{X}_N - p\right) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{\lfloor Nt \rfloor} \frac{X_i - \bar{X}_N}{\sqrt{p\,(1-p)}} + \frac{Nt - \lfloor Nt \rfloor}{\sqrt{N}\sqrt{p\,(1-p)}} \left(X_{\lfloor Nt \rfloor + 1} - \bar{X}_N\right) \qquad (3.14)
\end{aligned}$$

for $t \in (0,1)$.

Since the assumptions of Theorem A.4 are met for function $h(r) = \sup_t |r(t)|$, we have

$$\sup_{t \in (0,1)} |R_N(t)| \xrightarrow{d} \sup_{t \in (0,1)} |B(t)|, \quad \text{for } N \to \infty.$$

The last term in (3.14) is equal to 0 whenever $\lfloor Nt \rfloor$ is equal to an integer, so it can be easily shown that

$$\sup_{t \in (0,1)} |R_N(t)| = \max_{1 \le j \le N-1} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{j} \frac{X_i - \bar{X}_N}{\sqrt{p\,(1-p)}} \right|$$

and so

$$\max_{1\leq j\leq N-1}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{j}\frac{X_i-\bar{X}_N}{\sqrt{p\,(1-p)}}\right| \xrightarrow{d} \sup_{t\in(0,1)}|B(t)|, \quad \text{for } N\to\infty. \tag{3.15}$$

Link to the test statistic for known change point case

We can look at this test statistic as at generalization of the test statistic (3.4) from the Section 3.1. The numerator of the test statistic for known change point K can be rewritten in the following form:

$$\begin{aligned}
\hat{p}^{(K)}-\hat{q}^{(K)} &= \frac{1}{K}\sum_{i=1}^{K}X_i-\frac{1}{N-K}\sum_{i=K+1}^{N}X_i \\
&= \frac{1}{K}\sum_{i=1}^{K}X_i+\frac{1}{N-K}\sum_{i=1}^{K}X_i-\frac{1}{N-K}\sum_{i=K+1}^{N}X_i-\frac{1}{N-K}\sum_{i=1}^{K}X_i \\
&= \left(\frac{1}{K}+\frac{1}{N-K}\right)\sum_{i=1}^{K}-\frac{1}{N-K}\sum_{i=1}^{N}X_i \\
&= \frac{N}{K\,(N-K)}\sum_{i=1}^{K}\left(X_i-\bar{X}_N\right),
\end{aligned}$$

where $\bar{X}_N=\frac{1}{N}\sum_{i=1}^{N}X_i$. Using this form of numerator we can rewrite the test statistic in form

$$T^{(K)}=\sqrt{\frac{N^2}{K\,(N-K)}}\cdot\frac{1}{\sqrt{N}}\frac{\sum_{i=1}^{K}(X_i-\bar{X}_N)}{\sqrt{p\,(1-p)}}.$$

Now we can see that our test statistic (3.15) is closely related to the test statistic (3.4):

$$\max_{1\leq j\leq N-1}\left\{\sqrt{\frac{j\,(N-j)}{N^2}}\cdot\left|T^{(j)}\right|\right\}.$$

We could have used the test statistic $\max_{1\leq j\leq N-1}\left\{\left|T^{(j)}\right|\right\}$ for testing the hypothesis (3.1) with unknown change point as well, but this would lead us to extreme value distribution of the test statistic.

Distribution of test statistic under $\mathbf{H}_0$

Similarly to the section for known change point K, the real value of p is unknown and we need to estimate it. We use the estimator $\tilde{p}_N$ defined in (3.6). The estimator $\tilde{p}_N\,(1-\tilde{p}_N)$ is a consistent estimator of $p\,(1-p)$, so according to Cramér-Slutsky theorem (Theorem A.2) we can replace the $p\,(1-p)$ in (3.15) by this estimator while keeping the same asymptotic distribution:

$$T_{III}=\max_{1\leq j\leq N-1}\left|\frac{1}{\sqrt{N}}\frac{\sum_{i=1}^{j}(X_i-\bar{X}_N)}{\sqrt{\tilde{p}_N\,(1-\tilde{p}_N)}}\right| \xrightarrow{d} \sup_{t\in(0,1)}|B(t)|, \quad \text{for } N\to\infty \tag{3.16}$$

and we reject the null hypothesis (3.1) when

$$T_{III} > b_\alpha,$$

where b_α is such a value that $\mathsf{P}(\sup_{t\in(0,1)} |B(t)| \le b_\alpha) = 1-\alpha$. The table with these values for some choices of α can be found in Sen [1981, page 396]. The distribution of the supremum of Brownian bridge can be also expressed as an infinite sum:

$$\mathsf{P}\left(\sup_{t\in(0,1)} |B(t)| \le b\right) = 1 + 2\sum_{k=1}^{\infty}(-1)^k \mathrm{e}^{-2k^2b^2}, \quad b > 0. \tag{3.17}$$

More about this can be found in Billingsley [1968, page 85].

Distribution of test statistic under $\mathbf{H}_1$

Now let's look at the behaviour of the test statistic (3.16) under the alternative hypothesis. Then $p \neq q$ and there exists a change point $K_0 = \lfloor N\theta_0 \rfloor$ for some $\theta_0 \in (0,1)$. Let's look closer at the expression

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{K_0} \frac{X_i - \bar{X}_N}{\sqrt{\tilde{p}_N\,(1-\tilde{p}_N)}} \right| \tag{3.18}$$

which is the K_0-th element of expression over which we search for maximum in the test statistic (3.16).

With respect to the fact that

$$\left[\sum_{i=1}^{K_0} \left(X_i - \bar{X}_N \right) \right] \xrightarrow{\mathsf{P}} K_0\ (1-\kappa)\,(p-q)\,, \quad \text{for } N \to \infty,$$

and

$$\tilde{p}_N \xrightarrow{\mathsf{P}} \kappa\, p + (1-\kappa)\ q, \quad \text{for } N \to \infty,$$

we can approximate the expression (3.18) by

$$\left| \frac{K_0\,(1-\kappa)\,(p-q)}{\sqrt{N}\,\sqrt{(\kappa\, p + (1-\kappa)\, q)(1 - \kappa\, p - (1-\kappa)\, q)}} \right| = \frac{\lfloor N\theta_0 \rfloor}{\sqrt{N}}\, C \approx \sqrt{N}\,\theta_0\, C,$$

where C is a positive bounded number. Using this approximation we have

$$\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{K_0} \frac{X_i - \bar{X}_N}{\sqrt{\tilde{p}_N\,(1-\tilde{p}_N)}} \right| \xrightarrow{\mathsf{P}} +\infty, \quad \text{for } N \to \infty.$$

We have shown that one of the elements over which we search for maximum approaches $+\infty$ for $N \to \infty$ hence also the maximum needs to go to $+\infty$ for $N \to \infty$.

Test power

The process of approximation of the power for this case would be very complex and so it is not studied in this thesis.

3.1.3 Bootstrap and permutation test

The bootstrap algorithm and the procedure of permutation test for the Bernoulli case are described in this Section. The algorithms for both bootstrap and permutations and for all test statistics that were introduced in Sections 3.1.1 and 3.1.2 are similar, hence we describe only one of them and then explain the differences for the other cases.

The bootstrap algorithm for test statistic (3.8):

1. for $b = 1, \dots, B$, where B is the chosen number of bootstrap repetitions
   (a) sample N values from $X_1, \dots, X_N$ with replacement and denote $X^*_{1,b}, \dots, X^*_{N,b}$
   (b) compute he value of the test statistic $T_I^{(K)}$ for this sample and denote $T_{I,b}^{(K)*}$
2. find the $\lfloor (1-\alpha)B \rfloor$-th smallest value from $T_{I,1}^{(K)*}, \dots, T_{I,B}^{(K)*}$ and denote c_{boot}
3. the c_{boot} is the critical value for the test

In case of permutation test the only difference is in step 1a:

1a* sample N values from $X_1, \dots, X_N$ *without* replacement and denote $X^*_{1,b}, \dots, X^*_{N,b}$

Both the algorithms for bootstrap and permutation test are similar for the other test statistics (3.9) and (3.16).

The algorithm for simulating the power of these tests is again similar for all test statistics and both permutation and bootstrap tests, so only one version is described in detail.

The simulated power for bootstrap test with test statistic $T_I^{(K)}$ can be computed using this algorithm:

i. generate R data sets under the alternative hypothesis with particular values of p, q, N and K

ii. for each of these data sets compute the value of the test statistic $T_I^{(K)}$ and denote $T_I^{(K),r}$ for $r = 1, \dots, R$

iii. the simulated power is equal to portion of $T_I^{(K),r}$ for which $|T_I^{(K),r}| > c_{boot}$ where c_{boot} is critical value for the particular test

3.2 Multinomial distribution

In this section we generalize the results from the Section 3.1 for discrete distribution with more than two possible values — Multinomial distribution. Let s be the number of possible values (while $s > 2$). Let's consider $\boldsymbol{U}_1, \dots, \boldsymbol{U}_K$ to be independent random vectors distributed from multinomial distribution $\mathsf{M}(1; \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, \dots, p_s)^T$, $p_j \geq 0$ for $j = 1, \dots, s$ and $\sum_{j=1}^{s} p_j = 1$ and $\boldsymbol{U}_{K+1}, \dots, \boldsymbol{U}_N$ to be independent random vectors distributed from multinomial distribution $\mathsf{M}(1; \boldsymbol{q})$, where $\boldsymbol{q} = (q_1, \dots, q_s)^T$, $q_j \geq 0$ for $j = 1, \dots, s$ and

$\sum_{j=1}^{s} q_j = 1$. We can imagine each vector $\boldsymbol{U}_i = (U_{i,1}, \dots, U_{i,s})^T$ as one extract of a ball from a bag in which we have balls of s different colours — we extract only one ball (it means that only one element of vector $\boldsymbol{U}_i$ is non-zero) and it has a colour j with probability p_j, $j = 1, \dots, s$ (where the probability is equal to portion of the balls of colour j in the bag). We assume that $\mathsf{P}(U_{i,j} = 1) = p_j$ and $\mathsf{P}(U_{i,j} = 0) = 1 - p_j$ for $i = 1, \dots, K$ and $\mathsf{P}(U_{i,j} = 1) = q_j$ for $i = K+1, \dots, N$ and a change point K. Our hypothesis in this case has form

$$\begin{gathered} H_0 : \boldsymbol{p} = \boldsymbol{q} \\ \text{vs.} \\ H_1 : \boldsymbol{p} \neq \boldsymbol{q}. \end{gathered} \tag{3.19}$$

Similarly to the Bernoulli case we will start with known change point K (Section 3.2.1) and then extend the results for the case of the unknown change point (Section 3.2.2).

3.2.1 Known change point

In this Section we follow the technique described in Anděl [2005] which leads to a quadratic form test statistic.

Test statistic

Let's start with the maximum likelihood estimators of $\boldsymbol{p}$ and $\boldsymbol{q}$ for known change point K. The maximum likelihood estimator for $\boldsymbol{p} = (p_1, \dots, p_s)^T$ is $\hat{\boldsymbol{p}}^{(K)} = (\hat{p}_1^{(K)}, \dots, \hat{p}_s^{(K)})^T$ where

$$\hat{p}_j^{(K)} = \frac{\sum_{i=1}^{K} U_{i,j}}{K}$$

for $j = 1, \dots, s$.

Similarly, the maximum likelihood estimator for $\boldsymbol{q} = (q_1, \dots, q_s)^T$ is $\hat{\boldsymbol{q}}^{(K)} = (\hat{q}_1^{(K)}, \dots, \hat{q}_s^{(K)})^T$ where

$$\hat{q}_j^{(K)} = \frac{\sum_{i=K+1}^{N} U_{i,j}}{N - K}$$

for $j = 1, \dots, s$.

The vectors $\boldsymbol{U}_i$ for $i = 1, \dots, K$ are sampled from multinomial distribution $\mathsf{M}(1; \boldsymbol{p})$ so expectation of $\hat{p}_j^{(K)}$ is

$$\mathsf{E}\hat{p}_j^{(K)} = \mathsf{E}\left(\frac{1}{K}\sum_{i=1}^{K} U_{i,j}\right) = p_j,$$

for $j = 1, \dots, s$.

Similarly, the vectors $\boldsymbol{U}_i$ for $i = K+1, \dots, N$ are sampled from multinomial distribution $\mathsf{M}(1; \boldsymbol{q})$ so expectation of $\hat{q}_j^{(K)}$ is

$$\mathsf{E}\hat{q}_j^{(K)} = \mathsf{E}\left(\frac{1}{N-K}\sum_{i=K+1}^{N} U_{i,j}\right) = q_j,$$

for $j = 1, \ldots, s$.

The covariance matrix of $(\hat{p}_1^{(K)}, \ldots, \hat{p}_s^{(K)})^T$ has elements:

$$\operatorname{var} \hat{p}_j^{(K)} = \frac{1}{K^2} \operatorname{var} \left(\sum_{i=1}^{K} U_{i,j} \right) = \frac{p_j\,(1-p_j)}{K},$$

and

$$\begin{aligned} \operatorname{cov}(\hat{p}_j^{(K)}, \hat{p}_k^{(K)}) &= = \frac{1}{K^2} \mathsf{E} \left(\sum_{i=1}^{K} U_{i,j} \right) \left(\sum_{i=1}^{K} U_{i,k} \right) - p_j\,p_k = \\ &= \frac{1}{K^2} K\,(K-1)\,p_j\,p_k - p_j\,p_k = -\frac{p_j\,p_k}{K}, \end{aligned}$$

for $j, k = 1, \ldots, s$, $j \neq k$.

Similarly, for $(\hat{q}_1^{(K)}, \ldots, \hat{q}_s^{(K)})^T$ we have

$$\operatorname{var} \hat{q}_j^{(K)} = \frac{1}{(N-K)^2} \operatorname{var} \left(\sum_{i=K+1}^{N} U_{i,j} \right) = \frac{q_j\,(1-q_j)}{N-K},$$

and

$$\begin{aligned} \operatorname{cov}(\hat{q}_j^{(K)}, \hat{q}_k^{(K)}) &= \frac{1}{(N-K)^2} \mathsf{E} \left(\sum_{i=K+1}^{N} U_{i,j} \right) \left(\sum_{i=K+1}^{N} U_{i,k} \right) - q_j\,q_k = \\ &= \frac{1}{(N-K)^2} (N-K)\,(N-K-1)\,q_j\,q_k - q_j\,q_k \\ &= -\frac{q_j\,q_k}{N-K}, \end{aligned}$$

for $j, k = 1, \ldots, s$, $j \neq k$.

Let's look at the distribution of the differences $\hat{p}_j^{(K)} - \hat{q}_j^{(K)}$ under the null hypothesis now. It holds

$$\mathsf{E}_{H_0}(\hat{p}_j^{(K)} - \hat{q}_j^{(K)}) = \mathsf{E}\hat{p}_j^{(K)} - \mathsf{E}\hat{q}_j^{(K)} = p_j - p_j = 0,$$

$$\begin{aligned} \operatorname{var}_{H_0}(\hat{p}_j^{(K)} - \hat{q}_j^{(K)}) &= \operatorname{var} \hat{p}_j^{(K)} + \operatorname{var} \hat{q}_j^{(K)} = \\ &= \frac{p_j\,(1-p_j)}{K} + \frac{p_j\,(1-p_j)}{N-K} = \\ &= p_j\,(1-p_j) \left[\frac{1}{K} + \frac{1}{N-K} \right] = \\ &= \frac{N}{K(N-K)}\,p_j\,(1-p_j) \end{aligned}$$

for $j = 1, \ldots, s$ and

$$\begin{aligned} \operatorname{cov}_{H_0}(\hat{p}_j^{(K)} - \hat{q}_j^{(K)}, \hat{p}_k^{(K)} - \hat{q}_k^{(K)}) &= \operatorname{cov}(\hat{p}_j^{(K)}, \hat{p}_k^{(K)}) + \operatorname{cov}(\hat{q}_j^{(K)}, \hat{q}_k^{(K)}) = \\ &= -\frac{p_j\,p_k}{K} - \frac{p_j\,p_k}{N-K} = \\ &= -p_j\,p_k \left[\frac{1}{K} + \frac{1}{N-K} \right] = \\ &= -\frac{N}{K(N-K)}\,p_j\,p_k \end{aligned}$$

for $j,k=1,\dots,s,\ j\neq k$.

Denote $\boldsymbol{V}$ the variance matrix of the difference $\hat{\boldsymbol{p}}^{(K)}-\hat{\boldsymbol{q}}^{(K)}$ under the null hypothesis, and additionally denote

$$\boldsymbol{u}=\left(\sqrt{p_1},\dots,\sqrt{p_s}\right)^T,$$

$$\boldsymbol{A}=\boldsymbol{I}-\boldsymbol{u}\boldsymbol{u}^T=\begin{pmatrix} 1-p_1\ , & -\sqrt{p_1\,p_2}, & \cdots, & -\sqrt{p_1\,p_s} \\ -\sqrt{p_1\,p_2}, & 1-p_2\ , & \cdots, & -\sqrt{p_2\,p_s} \\ \cdots & \cdots & \cdots & \cdots \\ -\sqrt{p_1\,p_s}, & -\sqrt{p_2\,p_s}, & \cdots, & 1-p_s \end{pmatrix}$$

and

$$\boldsymbol{D}=\operatorname{Diag}\left\{\sqrt{\frac{Np_1}{K\,(N-K)}},\dots,\sqrt{\frac{Np_s}{K\,(N-K)}}\right\}.$$

Then it can be easily verified that

$$\boldsymbol{V}=\boldsymbol{DAD}.$$

The important fact for computing the rank of the variance matrix is that $\boldsymbol{u}^T\boldsymbol{u}=1$ and hence the matrix $\boldsymbol{A}$ is idempotent and for its rank we have

$$h(\boldsymbol{A})=\mathsf{Tr}\,\boldsymbol{A}=\mathsf{Tr}\,\boldsymbol{I}-\mathsf{Tr}\,\boldsymbol{u}\boldsymbol{u}^T=s-1.$$

The matrix $\boldsymbol{D}$ is nonsingular, hence by the multiplying the matrix $\boldsymbol{A}$ by $\boldsymbol{D}$ from right and left does not change the rank. Denote

$$\boldsymbol{V}^-=\boldsymbol{D}^{-2}=\operatorname{Diag}\left\{\frac{K\,(N-K)}{N\,p_1},\dots,\frac{K\,(N-K)}{N\,p_s}\right\}.$$

Using the idempotency of the matrix $\boldsymbol{A}$ we have

$$\boldsymbol{V}\boldsymbol{V}^-\boldsymbol{V}=\boldsymbol{DAD}\boldsymbol{D}^{-2}\boldsymbol{DAD}=\boldsymbol{D}\boldsymbol{A}^2\boldsymbol{D}=\boldsymbol{DAD}=\boldsymbol{V}.$$

Hence the matrix $\boldsymbol{V}^-$ is one of the pseudoinverses of the variance matrix $\boldsymbol{V}$ which has the rank of $s-1$.

Denote

$$\begin{aligned} Q^{(K)} &= (\hat{p}_1^{(K)}-\hat{q}_1^{(K)},\dots,\hat{p}_s^{(K)}-\hat{q}_s^{(K)})\boldsymbol{V}^-(\hat{p}_1^{(K)}-\hat{q}_1^{(K)},\dots,\hat{p}_s^{(K)}-\hat{q}_s^{(K)})^T \\ &= \frac{K\,(N-K)}{N}\sum_{j=1}^{s}\frac{(\hat{p}_j^{(K)}-\hat{q}_j^{(K)})^2}{p_j}. \end{aligned} \tag{3.20}$$

The difference of $\hat{p}_j^{(K)}-\hat{q}_j^{(K)}$ can be rephrased as

$$\begin{aligned} \hat{p}_j^{(K)}-\hat{q}_j^{(K)} &= \frac{1}{K}\sum_{i=1}^{K}U_{i,j}-\frac{1}{N-K}\sum_{i=K+1}^{N}U_{i,j}\pm\frac{1}{N-K}\sum_{i=1}^{K}U_{i,j} \\ &= \frac{N}{K\,(N-K)}\sum_{i=1}^{K}U_{i,j}-\frac{N}{N-K}\bar{U}_{N,j} \\ &= \frac{N}{K\,(N-K)}\sum_{i=1}^{K}\left(U_{i,j}-\bar{U}_{N,j}\right) \end{aligned}$$

and so we can rewrite the statistic (3.20) to the following form:

$$Q^{(K)} = \frac{N}{K\,(N-K)} \sum_{j=1}^{s} \left(\frac{\sum_{i=1}^{K} \left(U_{i,j} - \bar{U}_{N,j}\right)}{\sqrt{p_j}} \right)^2 . \tag{3.21}$$

Distribution of test statistic under $\mathbf{H}_0$

According to Theorem A.6 the statistic (3.21) has asymptotically χ^2_{s-1} distribution. However, the values of $p_j,\ j = 1,\ldots,s$ are unknown, hence we need to estimate them. Under H_0 the estimator $\tilde{p}_j = \frac{\sum_{i=1}^{N} U_{i,j}}{N}$ is a consistent estimator of $p_j,\ j = 1,\ldots,s$, and hence (according to Crámer-Slutsky Theorem — Theorem A.2) we can replace the p_j in (3.21) by this estimator while keeping the χ^2_{s-1} distribution:

$$Q_I^{(K)} = \frac{N}{K\,(N-K)} \sum_{j=1}^{s} \left(\frac{\sum_{i=1}^{K} \left(U_{i,j} - \bar{U}_{N,j}\right)}{\sqrt{\tilde{p}_j}} \right)^2 \overset{as.}{\sim} \chi^2_{s-1}. \tag{3.22}$$

Since the value of the test statistic (3.22) is large under the alternative hypothesis (see next paragraph), we reject the null hypothesis (3.19) on the asymptotical significance level α if

$$\left| Q_I^{(K)} \right| > \chi^2_{s-1}(1-\alpha), \tag{3.23}$$

where $\chi^2_{s-1}(1-\alpha)$ is $100(1-\alpha/2)\%$ quantile of the χ^2_{s-1}-distribution.

Distribution of test statistic under $\mathbf{H}_1$

Now let's look at the distribution of the test statistic (3.22) under the alternative hypothesis. It holds

$$\mathsf{E}\hat{p}_j^{(K)} = p_j,$$

$$\mathsf{E}\hat{q}_j^{(K)} = q_j \quad (q_j \neq p_j \text{ for at least one } j)$$

and

$$\tilde{p}_j \xrightarrow{\mathsf{P}} \kappa\, p_j + (1-\kappa)\, q_j, \quad \text{for } N \to \infty.$$

The term in sum on the left hand side in the expression (3.22) is then non-zero and the whole statistic $Q_I^{(K)}$ can be expressed as non-zero (positive) multiple of N which asymptotically approaches $+\infty$ for $N \to \infty$.

Test power

Studying of power function of the test (3.23) would lead us to approximation by non-central χ^2 distribution, but we will not pursue it here.

3.2.2 Unknown change point

Test statistic

In this section we introduce extension of test statistic (3.22) for unknown change point K.

Due to the condition $\sum_{j=1}^{s} p_j = 1$ the variance matrix $\boldsymbol{V}$ from previous section is singular and has a rank $s-1$. Let's denote $\boldsymbol{Z}_i = (U_{i,1}, \ldots, U_{i,s-1})^T$ (it means the vector $\boldsymbol{U}_i$ without the last component). The expectation of $\boldsymbol{Z}_i$ is equal to $(p_1, \ldots, p_{s-1})^T$ and the variance matrix $\boldsymbol{V}_Z$ with elements

$$\text{var } Z_{i,j} = p_j\,(1-p_j)$$

for $j = 1, \ldots, s-1$, and

$$\text{cov } (Z_{i,j}, Z_{i,k}) = -p_j\, p_k$$

for $j, k = 1, \ldots, s-1$, $j \neq k$, is nonsingular (positive definite).

The inverse of the matrix $\boldsymbol{V}_Z$ can be easily computed:

$$\boldsymbol{V}_Z^{-1} = \begin{pmatrix} \frac{p_1+p_s}{p_1\,p_s}, & \frac{1}{p_s}, & \ldots, & \frac{1}{p_s} \\ \frac{1}{p_s}, & \frac{p_2+p_s}{p_2\,p_s}, & \ldots, & \frac{1}{p_s} \\ \ldots & \ldots & \ldots & \ldots \\ \frac{1}{p_s}, & \frac{1}{p_s}, & \ldots, & \frac{p_{s-1}+p_s}{p_{s-1}\,p_s} \end{pmatrix}.$$

Let's define process $\{M_N(t),\, t \in (0,1)\}$ by

$$M_N(t) = \frac{1}{N} \left(\sum_{i=1}^{\lfloor Nt \rfloor} (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_N) \right)^T \boldsymbol{V}_Z^{-1} \left(\sum_{i=1}^{\lfloor Nt \rfloor} (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_N) \right)$$

for $t \in (0,1)$. The test statistic of our interest is

$$\begin{aligned} Q_{II} = \sup_{t\in(0,1)} M_N(t) &= \max_{1\le k\le N-1} \left\{ \frac{1}{N} \left(\sum_{i=1}^{k} (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_N) \right)^T \boldsymbol{V}_Z^{-1} \left(\sum_{i=1}^{k} (\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_N) \right) \right\} \\ &= \max_{1\le k\le N-1} \left\{ \frac{1}{N} \sum_{j=1}^{s} \left(\frac{\sum_{i=1}^{K} (U_{i,j} - \bar{U}_{N,j})}{\sqrt{p_j}} \right)^2 \right\}. \end{aligned} \tag{3.24}$$

Link to the test statistic for known change point case

Looking back to the test statistic for known change point (3.21) we can see that

$$Q_{II} = \max_{1\le j\le N-1} \left\{ \frac{j(N-j)}{N^2} \cdot Q^{(j)} \right\}.$$

Distribution of test statistic under $\mathbf{H}_0$

Let's look at the distribution of the test statistic (3.24).

The vectors $\boldsymbol{Z}_i$ are independent identically distributed with zero mean and nonsingular variance matrix so according to Theorem A.7

$$\{M_N(t),\, t \in (0,1)\} \xrightarrow{d} \{\sum_{j=1}^{s-1} B_j^2(t),\, t \in (0,1)\}, \quad \text{for } N \to \infty$$

and according to Theorem A.4 also

$$Q_{II} = \sup_{t\in(0,1)} |M_N(t)| \xrightarrow{d} \sup_{t\in(0,1)} \sum_{j=1}^{s-1} B_j^2(t), \quad \text{for } N \to \infty. \tag{3.25}$$

Similarly to the section for known change point K, the true values of $\boldsymbol{p}$ are unknown, so we need to estimate them. The estimators $\tilde{p}_j = \frac{\sum_{i=1}^N U_{i,j}}{N}$, $j = 1\ldots, s$ are consistent estimators of p_j, $j = 1, \ldots, s$ and hence (according to Crámer-Slutsky Theorem — Theorem A.2) we can replace the p_j, $j = 1, \ldots, s$ in (3.24) by these estimators while keeping the same distribution.

The value of the test statistic Q_{II} under the alternative hypothesis is large (see next paragraph) and so we reject the null hypothesis (3.19) if

$$Q_{II} > q_{1-\alpha}$$

where $q_{1-\alpha}$ is such a value that $\mathsf{P}(\sup_{t\in(0,1)} \sum_{j=1}^{s=1} B_j^2 \leq q_{1-\alpha}) = 1 - \alpha$. The table with these values for some choices of α and s can be found in Sen [1981, page 396].

Distribution of test statistic under $\mathbf{H}_1$

Under the alternative hypothesis there exists a change point $K_0 = \lfloor N\theta_0 \rfloor$ for some $\theta_0 \in (0,1)$. Let's look at the expression

$$\frac{1}{N}\sum_{j=1}^{s}\left(\frac{\sum_{i=1}^{K_0}(U_{i,j} - \bar{U}_{N,j})}{\sqrt{\tilde{p}_j}}\right)^2. \tag{3.26}$$

With respect to the fact that

$$\left[\sum_{i=1}^{K_0}\left(U_{i,j} - \bar{U}_{N,j}\right)\right] \xrightarrow{\mathsf{P}} K_0\,(1-\kappa)(p_j - q_j), \quad \text{for } N \to \infty$$

for $j = 1, \ldots, s$, and

$$\tilde{p}_j \xrightarrow{\mathsf{P}} \kappa p_j + (1-\kappa)q_j, \quad \text{for } N \to \infty,$$

for $j = 1, \ldots, s$, we can approximate the expression (3.26) by

$$\frac{1}{N}\sum_{j=1}^{s}\frac{\left(K_0\,(1-\kappa)(p_j - q_j)\right)^2}{\sqrt{\kappa p_j + (1-\kappa)q_j}} = \frac{K_0^2}{N}C \approx N\theta_0 C$$

where C is a positive bounded number. Using this approximation we can see, that

$$\frac{1}{N}\sum_{j=1}^{s}\left(\frac{\sum_{i=1}^{K_0}(U_{i,j}-\bar{U}_{N,j})}{\sqrt{\tilde{p}_j}}\right)^2 \xrightarrow{\mathsf{P}} +\infty, \quad \text{for } N\to\infty.$$

We have shown that one of the elements over which we search for maximum approaches $+\infty$ for $N \to \infty$ hence also the maximum itself needs to go to $+\infty$ for $N \to \infty$.

Test power

Similarly to the case of unknown change point in Bernoulli distribution case, the process of approximation of the power for this case would be very complex and so it is not studied in this thesis.

3.2.3 Bootstrap and permutation test

The bootstrap and permutation tests algorithm for Multinomial case are similar to those for Bernoulli case that are described in section 3.1.3. The power simulation for both bootstrap and permutation tests with both test statistics $Q_I^{(K)}$ and Q_{II} is following the algorithm described in section 3.1.3, too.

Chapter 4

Broken-line case simulations

Comparison of different approaches described in Chapter 2 has been done using simulations. The description of the simulations settings, implementation and results can be found in this chapter.

4.1 Settings and implementation

We consider 40 different settings for simulations. The values of β_0 and β_1 are common for all these settings and set to be $\beta_0 = +20$ and $\beta_1 = +5$. Four different values of β_2 are used — 0 (it means no change), -1, -5 and -10. We consider four different sample sizes N (30, 50, 100 and 500) and three different change points m ($\lceil N/2 \rceil$, $\lceil 2N/3 \rceil$ and $\lceil 3N/4 \rceil$). In case of $\beta_2 = 0$ (no change), we use only one "change point" equal to $\lceil N/2 \rceil$ because of possibility of using also the test for known change point. We assume the random errors ε_i, $i = 1, \ldots, N$ to be normally distributed. The true value of σ^2 is common for all the settings and equal to 1.

Hence in summary we have 4 data sets for "no change" and 12 data sets for each of other values β_2. We use the significance level $\alpha = 5$ %.

The test statistics (2.7) and (2.9) for known change point and the test statistic (2.13) for unknown change point have been implemented. For each of these statistics and each setting the permutation test with $B_P = 1\,000$ and $10\,000$ repetitions and the bootstrap with $B_B = 1\,000$ and $10\,000$ repetitions have been used.

4.2 Results

The results from "no change" case with known "change point" can be found in Tables 4.1 and 4.2, the unknown change point case in Table 4.3. Each of these tables contains the values of the critical value obtained from exact test (in case of test statistics (2.7) and (2.9)) or approximation gained from extreme value theory (in case of test statistic (2.13)) and critical values from permutation tests and bootstraps. For known change point test statistics we use the assumption of normality of the random errors for gaining the critical value from exact test.

The exact critical value for test statistic (2.7) is $z_{0.975} = 1.960$, for test statistic (2.9) it is $t_{27}(0.975) = 2.052$ for $N = 30$, $t_{47}(0.975) = 2.012$ for $N = 50$,

$t_{97}(0.975) = 1.985$ for $N = 100$ and $t_{497}(0.975) = 1.965$ for $N = 500$. The approximated critical value based on extreme value theory for test statistic (2.13) is equal to $c_{30} = 3.082$ for $N = 30$, $c_{50} = 3.089$ for $N = 50$, $c_{100} = 3.106$ for $N = 100$ and $c_{500} = 3.154$ for $N = 500$.

Similarly, other tables in this Section with the same structure contain the results for $\beta_2 = -1$ (Tables 4.4, 4.5 and 4.6), $\beta_2 = -5$ (Tables 4.7, 4.8 and 4.9) and $\beta_2 = -10$ (Tables 4.10, 4.11 and 4.12).

The power of the exact test (for known change point only) and simulated power of bootstrap and permutation tests (for both known and unknown change point) can be found in this Section, too. The original true values from the simulated data sets settings have been used for power calculation (it means that values of power in each table and each row are based on different values of $\boldsymbol{\beta}$, N and m). There is no power estimation for the tests based on data sets that have been simulated under the null hypothesis. The results for $\beta_2 = -1$ can be found in Tables 4.13, 4.14 and 4.15, for $\beta_2 = -5$ in Tables 4.16, 4.17 and 4.18, and for $\beta_2 = -10$ in Tables 4.19, 4.20 and 4.21.

Looking at the critical values for the test statistics (2.7) and (2.9) obtained by different methods, we can see that the critical values from the bootstrap and permutation test are close to the critical value from exact test already for small sample size ($N = 30$) and lower number of bootstrap and permutation repetitions ($B = 1\,000$) regardless of whether the original data were simulated under the null or alternative hypothesis. In case of the test statistic (2.13), the critical values obtained from bootstrap and permutation test are more or less similar to each other also already for small sample size and lower number of bootstrap and permutation repetitions and again regardless of whether the original data were simulated under the null or alternative hypothesis. However, the critical value from the extreme value theory approach is in most cases higher than those gained from bootstrap and permutations. It was mentioned in Section 2.2 that the approximations gained using extreme value theory usually converge very slowly. That is why we can see more conservative critical values coming from this approach than those coming from bootstrap and permutation test approximations.

The bootstrap and permutation test do not depend on the normality assumption of the random errors as the exact distributions of the test statistics (2.7) and (2.9) do. In case of omitting this assumption, the critical values coming from bootstrap and permutation tests would remain the same. The exact distributions of both test statistics are unknown in such case, but the asymptotic distributions of both test statistics were introduced in Chapter 2, so the critical values obtained from these asymptotic approaches could be used as approximations. It has been shown that the asymptotic distribution of both test statistics (2.7) and (2.9) is $\mathsf{N}(0, 1)$ and the particular critical value for our simulations would be 1.960.

The simulated power for known change point case from bootstrap and permutation tests is similar to the theoretical one computed from exact distribution for both test statistics (2.7) and (2.9). From Tables 4.13, 4.14 and 4.15 we can see that for small change in slope ($\beta_2 = -1$ in our simulated model) even 500 observations are not enough for gaining a test with solid power. On the other hand the biggest change that we simulated ($\beta_2 = -10$) is detectable with very high power already for small number of observations (see Tables 4.19, 4.20 and 4.21).

N	m	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	1.901	1.905	1.856	1.841
50	25	1.960	1.990	2.127	1.927	2.116
100	50	1.960	1.909	2.008	1.997	1.947
500	250	1.960	1.991	1.933	1.895	1.966

Table 4.1: The CRITICAL VALUES for test statistic $R_I^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, 0)$ (no change) and $\sigma^2 = 1$, and KNOWN "change point". The column $z_{1-\alpha/2}$ contains the critical value obtained from the exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	$\mathrm{t}_{N-3}(1-\alpha/2)$	perm1000	perm10000	boot1000	boot10000
30	15	2.052	2.094	2.033	2.019	2.076
50	25	2.012	2.099	1.973	1.925	2.020
100	50	1.985	2.007	1.974	1.931	1.986
500	250	1.965	1.887	1.958	1.945	1.969

Table 4.2: The CRITICAL VALUES for test statistic $R_{II}^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, 0)$ (no change) and $\sigma^2 = 1$, and KNOWN "change point". The column $\mathrm{t}_{N-3}(1-\alpha/2)$ contains the critical value obtained from the exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	EVT	perm1000	perm10000	boot1000	boot10000
30	15	3.082	2.841	2.829	2.805	2.800
50	25	3.089	2.881	2.929	2.877	2.931
100	50	3.106	2.855	2.849	2.898	2.837
500	250	3.154	2.858	2.820	2.820	2.834

Table 4.3: The CRITICAL VALUES for test statistic R_{III} and true values $\boldsymbol{\beta} = (20, 5, 0)$ (no change) and $\sigma^2 = 1$, and UNKNOWN "change point". The column EVT contains the critical value obtained from extreme value theory approximation, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	1.885	1.871	1.897	1.886
	20	1.960	1.740	1.727	1.695	1.726
	23	1.960	1.592	1.613	1.565	1.594
50	25	1.960	1.984	2.030	2.065	2.024
	34	1.960	1.893	1.894	1.940	1.866
	38	1.960	1.923	1.887	1.904	1.895
100	50	1.960	1.966	1.939	1.890	1.953
	67	1.960	1.960	1.934	1.967	1.971
	75	1.960	1.674	1.759	1.665	1.741
500	250	1.960	2.005	1.915	1.909	1.895
	334	1.960	1.925	1.926	2.083	1.923
	375	1.960	2.014	2.016	1.885	2.019

Table 4.4: The CRITICAL VALUES for test statistic $R_I^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -1)$ and $\sigma^2 = 1$, and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	$\mathrm{t}_{N-3}(1-\alpha/2)$	perm1000	perm10000	boot1000	boot10000
30	15	2.052	2.104	2.059	2.062	2.025
	20	2.052	1.932	2.035	1.972	2.030
	23	2.052	2.050	2.069	2.044	2.066
50	25	2.012	1.963	1.997	1.971	2.021
	34	2.012	1.915	2.039	1.987	1.998
	38	2.012	2.054	2.042	1.904	2.039
100	50	1.985	1.990	2.006	2.034	1.995
	67	1.985	1.927	1.952	1.907	1.994
	75	1.985	1.876	1.990	2.010	1.989
500	250	1.965	1.975	1.955	1.835	1.949
	334	1.965	1.961	1.975	2.012	1.982
	375	1.965	2.045	1.976	1.981	1.938

Table 4.5: The CRITICAL VALUES for test statistic $R_{II}^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -1)$ and $\sigma^2 = 1$, and KNOWN change point. The column $\mathrm{t}_{N-3}(1-\alpha/2)$ contains the critical value obtained from exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	EVT	perm1000	perm10000	boot1000	boot10000
30	15	3.082	2.333	2.426	2.399	2.421
	20	3.082	2.310	2.282	2.312	2.297
	23	3.082	2.136	2.122	2.235	2.156
50	25	3.089	2.716	2.749	2.701	2.755
	34	3.089	2.457	2.487	2.486	2.494
	38	3.089	2.583	2.531	2.451	2.524
100	50	3.106	2.688	2.689	2.620	2.690
	67	3.106	2.685	2.752	2.693	2.759
	75	3.106	2.381	2.391	2.423	2.393
500	250	3.154	2.670	2.783	2.820	2.824
	334	3.154	2.763	2.823	2.779	2.800
	375	3.154	2.919	2.913	2.888	2.929

Table 4.6: The CRITICAL VALUES for test statistic R_{III} and true values $\boldsymbol{\beta} = (20, 5, -1)$ and $\sigma^2 = 1$, and UNKNOWN change point. The column EVT contains the critical value obtained from extreme value theory approximation, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	2.267	2.296	2.215	2.294
	20	1.960	1.904	1.904	1.800	1.901
	23	1.960	1.808	1.875	1.854	1.828
50	25	1.960	1.820	1.832	1.763	1.826
	34	1.960	2.138	2.079	2.015	2.021
	38	1.960	1.866	2.004	1.922	1.976
100	50	1.960	2.023	1.996	1.981	1.969
	67	1.960	2.081	2.059	2.002	2.076
	75	1.960	1.790	1.773	1.831	1.754
500	250	1.960	1.999	2.094	2.111	2.083
	334	1.960	2.075	2.104	2.094	2.087
	375	1.960	2.050	1.951	1.898	1.950

Table 4.7: The CRITICAL VALUES for test statistic $R_I^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -5)$ and $\sigma^2 = 1$, and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	$t_{N-3}(1-\alpha/2)$	perm1000	perm10000	boot1000	boot10000
30	15	2.052	2.002	2.051	2.016	2.032
	20	2.052	1.972	2.031	2.080	2.049
	23	2.052	2.142	2.033	2.081	2.018
50	25	2.012	2.015	1.972	1.945	2.008
	34	2.012	2.157	2.015	2.027	2.035
	38	2.012	1.952	2.026	1.935	2.063
100	50	1.985	2.156	2.010	1.876	2.046
	67	1.985	2.038	1.981	2.109	1.991
	75	1.985	1.973	1.990	2.014	1.995
500	250	1.965	1.860	1.933	1.837	1.964
	334	1.965	1.957	1.960	1.919	1.940
	375	1.965	1.928	1.955	1.861	1.963

Table 4.8: The CRITICAL VALUES for test statistic $R_{II}^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -5)$ and $\sigma^2 = 1$, and KNOWN change point. The column $t_{N-3}(1-\alpha/2)$ contains the critical value obtained from exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	EVT	perm1000	perm10000	boot1000	boot10000
30	15	3.082	3.094	3.154	3.034	3.169
	20	3.082	2.536	2.462	2.447	2.478
	23	3.082	2.561	2.575	2.589	2.556
50	25	3.089	2.408	2.438	2.415	2.426
	34	3.089	2.735	2.716	2.650	2.688
	38	3.089	2.787	2.817	2.757	2.813
100	50	3.106	2.799	2.837	2.821	2.821
	67	3.106	2.804	2.837	2.966	2.843
	75	3.106	2.647	2.574	2.558	2.576
500	250	3.154	3.005	3.013	2.965	2.995
	334	3.154	3.114	3.066	3.015	3.086
	375	3.154	2.821	2.845	2.849	2.848

Table 4.9: The CRITICAL VALUES for test statistic R_{III} and true values $\boldsymbol{\beta} = (20, 5, -5)$ and $\sigma^2 = 1$, and UNKNOWN change point. The column EVT contains the critical value obtained from extreme value theory approximation, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	2.016	2.028	2.002	2.054
	20	1.960	2.156	2.168	2.109	2.141
	23	1.960	2.102	2.059	2.002	2.062
50	25	1.960	2.459	2.520	2.443	2.499
	34	1.960	2.282	2.258	2.247	2.257
	38	1.960	2.078	2.047	1.998	2.037
100	50	1.960	2.300	2.283	2.306	2.283
	67	1.960	2.239	2.190	2.141	2.172
	75	1.960	2.257	2.238	2.176	2.195
500	250	1.960	2.218	2.277	2.313	2.293
	334	1.960	2.277	2.383	2.501	2.429
	375	1.960	2.166	2.172	2.055	2.157

Table 4.10: The CRITICAL VALUES for test statistic $R_I^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -10)$ and $\sigma^2 = 1$, and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	$\mathrm{t}_{N-3}(1-\alpha/2)$	perm1000	perm10000	boot1000	boot10000
30	15	2.052	1.922	2.053	1.960	2.046
	20	2.052	2.003	2.012	1.979	2.065
	23	2.052	1.925	2.057	2.100	2.043
50	25	2.012	2.062	1.952	2.038	2.016
	34	2.012	1.922	1.979	2.062	2.026
	38	2.012	2.021	2.028	2.087	2.050
100	50	1.985	1.930	1.979	2.045	1.992
	67	1.985	2.027	1.996	1.906	2.006
	75	1.985	1.960	1.985	1.888	1.997
500	250	1.965	1.924	1.962	1.920	1.954
	334	1.965	1.889	1.976	2.136	1.968
	375	1.965	2.000	1.965	2.079	1.990

Table 4.11: The CRITICAL VALUES for test statistic $R_{II}^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -10)$ and $\sigma^2 = 1$, and KNOWN change point. The column $\mathrm{t}_{N-3}(1-\alpha/2)$ contains the critical value obtained from exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	EVT	perm1000	perm10000	boot1000	boot10000
30	15	3.082	2.721	2.773	2.748	2.764
	20	3.082	2.797	2.821	2.820	2.842
	23	3.082	2.730	2.696	2.826	2.739
50	25	3.089	3.362	3.357	3.223	3.331
	34	3.089	2.993	2.993	2.940	3.028
	38	3.089	2.841	2.829	2.805	2.800
100	50	3.106	3.075	3.094	3.071	3.116
	67	3.106	2.963	2.969	2.983	2.964
	75	3.106	3.027	3.011	2.929	3.045
500	250	3.154	3.342	3.290	3.304	3.320
	334	3.154	3.398	3.452	3.300	3.453
	375	3.154	3.039	3.072	3.111	3.048

Table 4.12: The CRITICAL VALUES for test statistic R_{III} and true values $\boldsymbol{\beta} = (20, 5, -10)$ and $\sigma^2 = 1$, and UNKNOWN change point. The column EVT contains the critical value obtained from extreme value theory approximation, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	exact	perm1000	perm10000	boot1000	boot10000
30	15	0.068	0.081	0.081	0.079	0.081
	20	0.064	0.110	0.114	0.122	0.114
	23	0.058	0.131	0.130	0.136	0.131
50	25	0.080	0.078	0.064	0.062	0.065
	34	0.071	0.081	0.081	0.074	0.085
	38	0.063	0.066	0.067	0.067	0.067
100	50	0.111	0.116	0.120	0.131	0.117
	67	0.093	0.108	0.115	0.107	0.107
	75	0.077	0.145	0.123	0.147	0.125
500	250	0.365	0.369	0.395	0.399	0.405
	334	0.272	0.301	0.301	0.253	0.301
	375	0.183	0.182	0.180	0.221	0.180

Table 4.13: The TEST POWER for test statistic $R_I^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -1)$ and $\sigma^2 = 1$, and KNOWN change point. The column exact contains the value for exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	exact	perm1000	perm10000	boot1000	boot10000
30	15	0.065	0.066	0.071	0.071	0.075
	20	0.064	0.089	0.069	0.080	0.070
	23	0.059	0.059	0.059	0.061	0.059
50	25	0.076	0.090	0.086	0.088	0.080
	34	0.069	0.078	0.060	0.067	0.064
	38	0.062	0.049	0.049	0.072	0.050
100	50	0.108	0.112	0.110	0.105	0.111
	67	0.090	0.114	0.107	0.120	0.100
	75	0.081	0.103	0.089	0.084	0.089
500	250	0.382	0.381	0.389	0.436	0.391
	334	0.276	0.288	0.284	0.274	0.282
	375	0.176	0.178	0.192	0.190	0.207

Table 4.14: The TEST POWER for test statistic $R_{II}^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -1)$ and $\sigma^2 = 1$, and KNOWN change point. The column exact contains the value for exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	perm1000	perm10000	boot1000	boot10000
30	15	0.113	0.088	0.097	0.088
	20	0.116	0.125	0.115	0.117
	23	0.171	0.179	0.133	0.161
50	25	0.073	0.067	0.075	0.067
	34	0.111	0.106	0.106	0.106
	38	0.083	0.090	0.107	0.091
100	50	0.100	0.100	0.116	0.100
	67	0.097	0.080	0.094	0.078
	75	0.169	0.165	0.157	0.165
500	250	0.283	0.240	0.228	0.226
	334	0.182	0.166	0.181	0.174
	375	0.089	0.090	0.093	0.088

Table 4.15: The TEST POWER for test statistic R_{III} and true values $\boldsymbol{\beta} = (20, 5, -1)$ and $\sigma^2 = 1$, and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	exact	perm1000	perm10000	boot1000	boot10000
30	15	0.506	0.416	0.402	0.439	0.402
	20	0.404	0.446	0.446	0.487	0.448
	23	0.250	0.313	0.285	0.297	0.305
50	25	0.723	0.771	0.771	0.791	0.771
	34	0.565	0.499	0.532	0.555	0.555
	38	0.381	0.412	0.361	0.390	0.369
100	50	0.950	0.932	0.937	0.939	0.939
	67	0.859	0.821	0.828	0.843	0.823
	75	0.667	0.720	0.727	0.702	0.732
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000	1.000

Table 4.16: The TEST POWER for test statistic $R_I^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -5)$ and $\sigma^2 = 1$, and KNOWN change point. The column exact contains the value for exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	exact	perm1000	perm10000	boot1000	boot10000
30	15	0.375	0.521	0.507	0.516	0.510
	20	0.450	0.434	0.411	0.398	0.404
	23	0.273	0.219	0.246	0.237	0.250
50	25	0.761	0.702	0.722	0.731	0.706
	34	0.503	0.495	0.546	0.542	0.540
	38	0.334	0.395	0.365	0.399	0.351
100	50	0.940	0.914	0.930	0.947	0.926
	67	0.854	0.838	0.844	0.815	0.843
	75	0.777	0.655	0.647	0.635	0.646
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000	1.000

Table 4.17: The TEST POWER for test statistic $R_{II}^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -5)$ and $\sigma^2 = 1$, and KNOWN change point. The column exact contains the value for exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	perm1000	perm10000	boot1000	boot10000
30	15	0.238	0.223	0.252	0.216
	20	0.364	0.387	0.390	0.379
	23	0.223	0.222	0.218	0.225
50	25	0.699	0.689	0.696	0.691
	34	0.396	0.407	0.435	0.422
	38	0.216	0.208	0.232	0.208
100	50	0.846	0.837	0.840	0.840
	67	0.686	0.675	0.622	0.674
	75	0.537	0.563	0.572	0.562
500	250	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000
	375	0.999	0.998	0.998	0.998

Table 4.18: The TEST POWER for test statistic R_{III} and true values $\boldsymbol{\beta} = (20, 5, -5)$ and $\sigma^2 = 1$, and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	exact	perm1000	perm10000	boot1000	boot10000
30	15	0.977	0.983	0.982	0.983	0.980
	20	0.930	0.909	0.908	0.918	0.911
	23	0.728	0.703	0.710	0.733	0.710
50	25	0.999	0.994	0.993	0.994	0.994
	34	0.989	0.973	0.974	0.974	0.974
	38	0.912	0.898	0.901	0.907	0.901
100	50	1.000	1.000	1.000	1.000	1.000
	67	1.000	1.000	1.000	1.000	1.000
	75	0.998	0.991	0.991	0.992	0.992
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000	1.000

Table 4.19: The TEST POWER for test statistic $R_I^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -10)$ and $\sigma^2 = 1$, and KNOWN change point. The column exact contains the value for exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	exact	perm1000	perm10000	boot1000	boot10000
30	15	0.968	0.987	0.977	0.985	0.978
	20	0.934	0.929	0.927	0.931	0.915
	23	0.729	0.760	0.720	0.708	0.725
50	25	0.994	0.998	0.999	0.998	0.998
	34	0.962	0.984	0.983	0.983	0.983
	38	0.923	0.904	0.902	0.888	0.897
100	50	1.000	1.000	1.000	1.000	1.000
	67	1.000	1.000	1.000	1.000	1.000
	75	0.996	0.995	0.994	0.996	0.994
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000	1.000

Table 4.20: The TEST POWER for test statistic $R_{II}^{(m)}$ and true values $\boldsymbol{\beta} = (20, 5, -10)$ and $\sigma^2 = 1$, and KNOWN change point. The column exact contains the value for exact test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	m	perm1000	perm10000	boot1000	boot10000
30	15	0.937	0.928	0.931	0.929
	20	0.815	0.811	0.811	0.804
	23	0.571	0.578	0.538	0.565
50	25	0.966	0.966	0.973	0.968
	34	0.927	0.927	0.933	0.920
	38	0.774	0.781	0.790	0.791
100	50	1.000	1.000	1.000	1.000
	67	1.000	1.000	0.999	1.000
	75	0.967	0.968	0.973	0.967
500	250	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000

Table 4.21: The TEST POWER for test statistic R_{III} and true values $\boldsymbol{\beta} = (20, 5, -10)$ and $\sigma^2 = 1$, and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

Chapter 5

Discrete case simulations

The different approaches described in Chapter 3 have been compared using simulations. The description of the simulations settings, implementation and results for the Bernoulli distribution case can be found in Section 5.1 and for the Multinomial distribution in Section 5.2.

5.1 Bernoulli distribution

5.1.1 Settings and implementation

We consider 40 different settings for the Bernoulli case data. The true value of p is common for all these settings and set to be 0.5. Four different values of q are considered — 0.5 (it means no change), 0.6, 0.75 and 0.9. We consider four different sample sizes N (30, 50, 100 and 500) and three different change points K ($\lceil N/2 \rceil$, $\lceil 2N/3 \rceil$ and $\lceil 3N/4 \rceil$). In case of $q = 0.5$ (no change), we use only one "change point" equal to $\lceil N/2 \rceil$ because of possibility of using also the tests for known change point.

Hence in summary we have 4 data sets for "no change" and 12 data sets for each of other values q. We use the significance level $\alpha = 5$ %.

The test statistics (3.8) and (3.9) for known change point and the test statistic (3.16) for unknown change point have been implemented. For each of these statistics and each setting the permutation test with $B_P = 1\,000$ and $10\,000$ repetitions and the bootstrap with $B_B = 1\,000$ and $10\,000$ repetitions have been used.

5.1.2 Results

The results from "no change" case with known "change point" can be found in Tables 5.1 and 5.2, the unknown change point case in Table 5.3. Each of these tables contains the values of the critical value obtained from asymptotic behaviour of that test statistic and critical values from permutation tests and bootstraps.

The asymptotic critical value for test statistics (3.8) and (3.9) is $z_{0.975} = 1.960$ and the one for test statistic (3.16) is computed from (3.17) and equal to $b_{0.025} = 1.358$.

Similarly, other tables in this Section with the same structure contain the results for $q = 0.6$ (Tables 5.4, 5.5 and 5.6), $q = 0.75$ (Tables 5.7, 5.8 and 5.9) and $q = 0.9$ (Tables 5.10, 5.11 and 5.12).

The approximated power for the asymptotic tests (for known change point only) and simulated power for bootstrap and permutation tests (for both known and unknown change point) can be found in this Section, too. The original true values from the simulated data sets settings have been used for power calculation (it means that values of power in each table and each row are based on different values of q, N and K). There is no power estimation for the tests based on data sets that have been simulated under null hypothesis. The results for $q = 0.6$ can be found in Tables 5.13, 5.14 and 5.15, for $q = 0.75$ in Tables 5.16, 5.17 and 5.18, and for $q = 0.9$ in Tables 5.19, 5.20 and 5.21.

The critical values obtained from bootstrap and permutation test are similar to the one obtained from asymptotic test in case of known change point regardless of whether the original data were simulated under the null or alternative hypothesis (and for both test statistics (3.8) and (3.9)). The approximated values vary a bit for small sample sizes, but they are quite stable for $N = 500$. In case of unknown change point the critical values from bootstrap and permutation tests are comparable to the one obtained from asymptotic test for $N = 500$. For smaller sample sizes the critical value from asymptotic test is higher than those gained from bootstrap and permutation tests — this could suggest slower convergence of the asymptotical approach.

The approximated power coming from the asymptotic approach is higher for test using test statistic (2.9) than for the one using test statistic (2.7) in all studied cases. However, this is not a general property. For some combinations of true values of p, q and κ $(= \frac{K}{N})$ the approximated power is higher for (2.7) than for (2.9). In case of small difference of p and q (the simulated case with $p = 0.5$ and $q = 0.6$) the power is not very high for any of the tests even for $N = 500$. The middle-size case change ($q = 0.75$) is detectable with power higher than 90 % only for $N = 500$ case. Even for the biggest studied change case ($q = 0.9$) the sample size of $N = 30$ is quite poor for unknown change point if the real K is far from $N/2$ (see Table 5.21).

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	2.196	2.196	1.894	1.894
50	25	1.960	1.698	1.698	1.770	1.986
100	50	1.960	2.026	2.026	2.000	2.002
500	250	1.960	1.888	1.879	1.968	1.968

Table 5.1: The CRITICAL VALUES for test statistic $T_I^{(K)}$ and true values $p = 0.5$ and $q = 0.5$ (no change), and KNOWN "change point". The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	2.397	2.397	2.019	2.019
50	25	1.960	1.750	1.750	2.070	2.070
100	50	1.960	2.069	2.069	2.041	1.952
500	250	1.960	1.885	1.885	1.984	1.977

Table 5.2: The CRITICAL VALUES for test statistic $T_{II}^{(K)}$ and true values $p = 0.5$ and $q = 0.5$ (no change), and KNOWN "change point". The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$b_{\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	none	1.358	1.293	1.269	1.278	1.278
50	none	1.358	1.279	1.302	1.280	1.279
100	none	1.358	1.338	1.313	1.305	1.306
500	none	1.358	1.336	1.338	1.366	1.338

Table 5.3: The CRITICAL VALUES for test statistic T_{III} and true values $p = 0.5$ and $q = 0.5$ (no change), and UNKNOWN "change point". The column $b_{\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	2.196	2.196	1.894	1.894
	20	1.960	1.917	1.917	1.936	1.936
	23	1.960	1.961	1.961	1.961	1.961
50	25	1.960	1.986	1.986	1.986	1.986
	34	1.960	2.044	2.044	2.015	2.015
	38	1.960	2.147	1.829	1.893	1.986
100	50	1.960	1.830	1.830	2.000	2.002
	67	1.960	2.066	2.066	1.959	1.959
	75	1.960	1.967	1.967	1.973	1.963
500	250	1.960	1.973	1.973	1.977	1.971
	334	1.960	1.903	1.903	1.839	1.962
	375	1.960	2.067	2.067	1.914	1.963

Table 5.4: The CRITICAL VALUES for test statistic $T_I^{(K)}$ and true values $p = 0.5$ and $q = 0.6$ and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	2.397	2.397	2.019	2.019
	20	1.960	2.394	2.394	2.236	2.236
	23	1.960	2.516	2.516	2.340	2.516
50	25	1.960	2.070	2.070	2.070	2.092
	34	1.960	2.242	2.242	2.138	2.138
	38	1.960	2.034	2.034	2.036	2.216
100	50	1.960	2.295	1.861	2.085	2.025
	67	1.960	1.883	2.090	2.089	2.077
	75	1.960	2.032	2.032	2.072	2.072
500	250	1.960	1.981	1.981	1.895	1.979
	334	1.960	2.098	1.918	2.030	1.978
	375	1.960	1.883	1.883	2.029	1.975

Table 5.5: The CRITICAL VALUES for test statistic $T_{II}^{(K)}$ and true values $p = 0.5$ and $q = 0.6$ and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$b_{\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.358	1.293	1.269	1.279	1.278
	20	1.358	1.291	1.291	1.275	1.275
	23	1.358	1.269	1.269	1.291	1.278
50	25	1.358	1.278	1.294	1.276	1.280
	34	1.358	1.294	1.294	1.301	1.279
	38	1.358	1.257	1.291	1.273	1.291
100	50	1.358	1.306	1.326	1.339	1.316
	67	1.358	1.286	1.313	1.305	1.312
	75	1.358	1.278	1.304	1.293	1.301
500	250	1.358	1.328	1.321	1.294	1.336
	334	1.358	1.396	1.337	1.320	1.327
	375	1.358	1.328	1.325	1.358	1.336

Table 5.6: The CRITICAL VALUES for test statistic T_{III} and true values $p = 0.5$ and $q = 0.6$ and UNKNOWN change point. The column $b_{\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	1.826	1.826	2.148	1.826
	20	1.960	1.811	2.070	2.070	2.070
	23	1.960	1.771	1.771	1.961	1.961
50	25	1.960	1.732	1.732	2.040	2.006
	34	1.960	1.857	1.857	1.879	2.015
	38	1.960	1.814	1.814	2.040	1.987
100	50	1.960	2.041	2.041	2.007	2.006
	67	1.960	1.933	1.959	1.933	1.959
	75	1.960	2.085	2.085	1.967	1.963
500	250	1.960	2.045	2.045	1.950	1.929
	334	1.960	1.970	1.966	1.920	1.933
	375	1.960	1.966	1.966	1.866	1.971

Table 5.7: The CRITICAL VALUES for test statistic $T_I^{(K)}$ and true values $p = 0.5$ and $q = 0.75$ and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	1.936	1.936	2.335	1.936
	20	1.960	1.945	2.390	2.390	2.254
	23	1.960	2.251	2.251	2.340	2.516
50	25	1.960	1.786	1.786	2.092	2.092
	34	1.960	2.025	2.025	2.025	2.138
	38	1.960	1.900	1.900	2.088	2.216
100	50	1.960	2.085	2.085	1.921	2.057
	67	1.960	2.077	2.077	2.118	2.031
	75	1.960	2.154	2.154	1.910	2.068
500	250	1.960	2.053	2.053	1.957	1.982
	334	1.960	1.873	1.965	1.965	1.966
	375	1.960	1.971	1.998	1.946	1.998

Table 5.8: The CRITICAL VALUES for test statistic $T_{II}^{(K)}$ and true values $p = 0.5$ and $q = 0.75$ and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$b_{\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.358	1.278	1.278	1.278	1.252
	20	1.358	1.269	1.269	1.278	1.278
	23	1.358	1.290	1.290	1.267	1.277
50	25	1.358	1.270	1.328	1.259	1.273
	34	1.358	1.276	1.299	1.250	1.273
	38	1.358	1.299	1.288	1.336	1.273
100	50	1.358	1.306	1.306	1.286	1.317
	67	1.358	1.265	1.332	1.321	1.311
	75	1.358	1.308	1.308	1.304	1.300
500	250	1.358	1.285	1.320	1.366	1.331
	334	1.358	1.317	1.330	1.344	1.329
	375	1.358	1.363	1.327	1.345	1.330

Table 5.9: The CRITICAL VALUES for test statistic T_{III} and true values $p = 0.5$ and $q = 0.75$ and UNKNOWN change point. The column $b_{\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	1.992	1.992	1.894	1.992
	20	1.960	1.917	1.917	1.936	1.936
	23	1.960	2.158	2.158	1.961	1.961
50	25	1.960	2.160	2.160	1.987	1.987
	34	1.960	2.046	2.046	1.926	1.966
	38	1.960	2.041	2.041	1.879	1.947
100	50	1.960	2.144	2.144	1.914	1.946
	67	1.960	1.823	2.084	1.984	1.959
	75	1.960	1.925	1.925	2.013	1.980
500	250	1.960	1.983	1.983	2.040	1.973
	334	1.960	2.074	1.902	2.041	1.946
	375	1.960	1.843	1.948	1.891	1.979

Table 5.10: The CRITICAL VALUES for test statistic $T_I^{(K)}$ and true values $p = 0.5$ and $q = 0.9$ and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$z_{1-\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.960	2.138	2.138	2.138	2.138
	20	1.960	2.394	2.394	2.329	2.254
	23	1.960	2.791	2.791	2.516	2.516
50	25	1.960	2.269	2.269	1.961	2.070
	34	1.960	2.050	2.050	2.192	2.076
	38	1.960	2.011	2.011	2.334	2.255
100	50	1.960	2.195	2.195	1.984	2.024
	67	1.960	1.927	2.089	2.110	1.982
	75	1.960	2.174	2.174	1.902	2.075
500	250	1.960	1.991	1.991	1.968	1.967
	334	1.960	1.922	1.922	2.066	1.978
	375	1.960	1.823	2.007	1.935	1.988

Table 5.11: The CRITICAL VALUES for test statistic $T_{II}^{(K)}$ and true values $p = 0.5$ and $q = 0.9$ and KNOWN change point. The column $z_{1-\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$b_{\alpha/2}$	perm1000	perm10000	boot1000	boot10000
30	15	1.358	1.275	1.275	1.275	1.278
	20	1.358	1.291	1.291	1.252	1.267
	23	1.358	1.278	1.278	1.290	1.278
50	25	1.358	1.327	1.296	1.310	1.285
	34	1.358	1.285	1.296	1.266	1.300
	38	1.358	1.260	1.278	1.273	1.285
100	50	1.358	1.278	1.312	1.331	1.325
	67	1.358	1.347	1.306	1.267	1.321
	75	1.358	1.317	1.325	1.278	1.313
500	250	1.358	1.325	1.333	1.279	1.340
	334	1.358	1.366	1.366	1.323	1.330
	375	1.358	1.267	1.335	1.386	1.328

Table 5.12: The CRITICAL VALUES for test statistic T_{III} and true values $p = 0.5$ and $q = 0.9$ and UNKNOWN change point. The column $b_{\alpha/2}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	asympt	perm1000	perm10000	boot1000	boot10000
30	15	0.084	0.048	0.048	0.087	0.087
	20	0.078	0.075	0.075	0.075	0.075
	23	0.071	0.057	0.057	0.057	0.057
50	25	0.108	0.107	0.107	0.127	0.107
	34	0.098	0.068	0.068	0.082	0.082
	38	0.089	0.061	0.100	0.097	0.088
100	50	0.170	0.193	0.193	0.175	0.173
	67	0.153	0.136	0.136	0.155	0.155
	75	0.136	0.131	0.131	0.127	0.137
500	250	0.614	0.630	0.630	0.617	0.633
	334	0.561	0.595	0.595	0.618	0.578
	375	0.492	0.461	0.461	0.515	0.502

Table 5.13: The TEST POWER for test statistic $T_I^{(K)}$ and true values $p = 0.5$ and $q = 0.6$ and KNOWN change point. The column asympt contains the value for asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	asympt	perm1000	perm10000	boot1000	boot10000
30	15	0.086	0.048	0.048	0.070	0.070
	20	0.082	0.045	0.045	0.077	0.077
	23	0.076	0.057	0.057	0.077	0.062
50	25	0.110	0.107	0.107	0.127	0.091
	34	0.103	0.067	0.067	0.081	0.081
	38	0.094	0.105	0.105	0.105	0.083
100	50	0.173	0.100	0.193	0.131	0.177
	67	0.159	0.198	0.155	0.155	0.155
	75	0.142	0.144	0.144	0.131	0.131
500	250	0.617	0.630	0.630	0.652	0.633
	334	0.569	0.527	0.595	0.548	0.573
	375	0.503	0.532	0.532	0.494	0.507

Table 5.14: The TEST POWER for test statistic $T_{II}^{(K)}$ and true values $p = 0.5$ and $q = 0.6$ and KNOWN change point. The column asympt contains the value for asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.043	0.066	0.051	0.051
	20	0.040	0.040	0.057	0.057
	23	0.060	0.060	0.047	0.051
50	25	0.105	0.101	0.106	0.103
	34	0.081	0.081	0.079	0.085
	38	0.085	0.069	0.074	0.069
100	50	0.141	0.128	0.118	0.131
	67	0.125	0.108	0.116	0.109
	75	0.105	0.089	0.098	0.090
500	250	0.543	0.548	0.565	0.534
	334	0.373	0.432	0.450	0.440
	375	0.311	0.315	0.290	0.303

Table 5.15: The TEST POWER for test statistic T_{III} and true values $p = 0.5$ and $q = 0.6$ and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	asympt	perm1000	perm10000	boot1000	boot10000
30	15	0.286	0.394	0.394	0.249	0.394
	20	0.241	0.344	0.215	0.215	0.215
	23	0.191	0.246	0.246	0.173	0.173
50	25	0.445	0.555	0.555	0.414	0.437
	34	0.377	0.397	0.397	0.397	0.372
	38	0.314	0.386	0.386	0.265	0.300
100	50	0.740	0.695	0.695	0.718	0.718
	67	0.677	0.678	0.660	0.678	0.660
	75	0.597	0.547	0.547	0.601	0.602
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	0.999	0.999	0.999	0.999	0.999

Table 5.16: The TEST POWER for test statistic $T_I^{(K)}$ and true values $p = 0.5$ and $q = 0.75$ and KNOWN change point. The column asympt contains the value for asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	asympt	perm1000	perm10000	boot1000	boot10000
30	15	0.310	0.394	0.394	0.249	0.394
	20	0.293	0.366	0.237	0.271	0.271
	23	0.251	0.242	0.242	0.238	0.181
50	25	0.472	0.555	0.555	0.437	0.437
	34	0.441	0.405	0.405	0.405	0.400
	38	0.389	0.401	0.401	0.353	0.329
100	50	0.762	0.695	0.695	0.735	0.710
	67	0.731	0.661	0.661	0.657	0.670
	75	0.671	0.592	0.592	0.661	0.621
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	1.000	0.999	0.999	0.999	0.999

Table 5.17: The TEST POWER for test statistic $T_{II}^{(K)}$ and true values $p = 0.5$ and $q = 0.75$ and KNOWN change point. The column asympt contains the value for asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.249	0.249	0.249	0.274
	20	0.168	0.168	0.153	0.153
	23	0.085	0.085	0.107	0.103
50	25	0.393	0.356	0.406	0.389
	34	0.279	0.263	0.310	0.281
	38	0.190	0.198	0.165	0.203
100	50	0.639	0.639	0.654	0.629
	67	0.590	0.525	0.535	0.540
	75	0.408	0.408	0.417	0.423
500	250	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000
	375	0.988	0.993	0.990	0.993

Table 5.18: The TEST POWER for test statistic T_{III} and true values $p = 0.5$ and $q = 0.75$ and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	asympt	perm1000	perm10000	boot1000	boot10000
30	15	0.684	0.675	0.675	0.733	0.675
	20	0.592	0.641	0.641	0.629	0.629
	23	0.460	0.377	0.377	0.541	0.541
50	25	0.895	0.864	0.864	0.906	0.906
	34	0.839	0.825	0.825	0.858	0.858
	38	0.754	0.736	0.736	0.817	0.791
100	50	0.996	0.994	0.994	0.997	0.997
	67	0.993	0.993	0.982	0.984	0.984
	75	0.984	0.974	0.974	0.970	0.972
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000	1.000

Table 5.19: The TEST POWER for test statistic $T_I^{(K)}$ and true values $p = 0.5$ and $q = 0.9$ and KNOWN change point. The column asympt contains the value for asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	asympt	perm1000	perm10000	boot1000	boot10000
30	15	0.757	0.675	0.675	0.675	0.675
	20	0.779	0.625	0.625	0.701	0.701
	23	0.738	0.531	0.531	0.572	0.572
50	25	0.929	0.864	0.864	0.925	0.906
	34	0.940	0.894	0.894	0.859	0.894
	38	0.921	0.876	0.876	0.797	0.831
100	50	0.998	0.994	0.994	0.997	0.997
	67	0.999	0.993	0.985	0.985	0.991
	75	0.998	0.976	0.976	0.980	0.976
500	250	1.000	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000	1.000

Table 5.20: The TEST POWER for test statistic $T_{II}^{(K)}$ and true values $p = 0.5$ and $q = 0.9$ and KNOWN change point. The column asympt contains the value for asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.629	0.629	0.629	0.623
	20	0.428	0.428	0.502	0.483
	23	0.223	0.223	0.219	0.223
50	25	0.834	0.853	0.847	0.858
	34	0.710	0.706	0.730	0.700
	38	0.538	0.502	0.503	0.490
100	50	0.989	0.989	0.986	0.988
	67	0.947	0.960	0.967	0.955
	75	0.879	0.876	0.895	0.880
500	250	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000

Table 5.21: The TEST POWER for test statistic T_{III} and true values $p = 0.5$ and $q = 0.9$ and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

5.2 Mutlinomial distribution

5.2.1 Settings and implementation

Similarly to the previous Section, we consider 40 different settings for the Multinomial case data as well. The number of possible values is the same for all settings and set to be 4 (it means $s = 4$ and the set of possible values is set to be $\{1, 2, 3, 4\}$). The vector of true values of $\boldsymbol{p} = (p_1, \ldots, p_s)^T$ is common for all settings, too, and set to be $(0.25, 0.25, 0.25, 0.25)^T$. Four different vectors of values for $\boldsymbol{q} = (q_1, \ldots, q_s)^T$ are considered — no change $(0.25, 0.25, 0.25, 0.25)^T$, quite small change in only two parameters $(0.1, 0.4, 0.25, 0.25)^T$, quite small change in all parameters $(0.1, 0.2, 0.3, 0.4)^T$ and bigger change in all parameters $(0.1, 0.1, 0.1, 0.7)^T$. The considered values for sample size and change point are the same as in Bernoulli case ($N = 30, 50, 100, 500$ and $K = \lceil N/2 \rceil, \lceil 2N/3 \rceil, \lceil 3N/4 \rceil$). However, for the case of unknown change point only $N = 30, 50, 100$ are used (case of $N = 500$ is already quite computationally intensive and time-consuming in this case). In case of no change we again use only one "change point" equal to $\lceil N/2 \rceil$ because of possibility of using also the test for known change point. We use the significance level $\alpha = 5$ %.

The test statistic (3.22) for known change point and the test statistic (3.25) for unknown change point have been implemented. For each of these statistics and each setting the permutation test with $B_P = 1\,000$ and $10\,000$ repetitions and the bootstrap with $B_B = 1\,000$ and $10\,000$ repetitions have been used.

5.2.2 Results

The results from "no change" case with known "change point" for multinomial data case can be found in Table 5.22, the unknown change point case in Table 5.23. Each of these tables contains the critical value obtained from asymptotic behaviour of the particular test statistic and critical values from permutation tests and bootstraps.

The asymptotic critical value for test statistic (3.22) is $\chi^2_3(0.95) = 7.815$ and the one for test statistic (3.25) is equal to $q_{0.95} = 3.056$.

Similarly, other tables with the same structure contain the results for $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ (Tables 5.24 and 5.25), $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ (Tables 5.26 and 5.27) and $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ (Tables 5.28 and 5.29).

The simulated power for bootstrap and permutation tests can be found in this Section, too. The original true values from the simulated data sets settings have been used for power calculation (it means that values of power in each table and each row are based on different values of $\boldsymbol{q}$, N and K). There is no power estimation for the tests based on data sets that have been simulated under the null hypothesis. The results for $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ can be found in Tables 5.30 and 5.31, for $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ in Tables 5.32 and 5.33, and for $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ in Tables 5.34 and 5.35.

The critical values obtained from bootstrap and permutation test are similar to the one obtained from asymptotic test in case of known change point regardless of whether the original data were simulated under the null or alternative hypothesis. In case of unknown change point the critical value obtained from asymptotic test is higher than those obtained from bootstrap and permutation tests — this could suggest slower convergence of the asymptotic approach.

The approximated power for known change point is similar for cases of $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ and it is quite high ($> 90\,\%$) only for $N = 500$. In case of $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ the sample size of 100 is already enough for detecting the change with power higher than 90 %. Similarly, for the unknown change point case the approximated power for $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ is similar to each other and not very high even for $N = 100$. On the other side, in case of $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ the approximated power exceeds 90 % for $N = 100$ if change point is not too close to the edge of the sample (beginning or end).

N	K	$\chi^2_3(1-\alpha)$	perm1000	perm10000	boot1000	boot10000
30	15	7.815	7.610	7.810	7.035	7.505
50	25	7.815	7.510	7.703	7.853	7.535
100	50	7.815	8.016	7.860	7.703	7.759
500	250	7.815	7.641	7.943	7.974	7.815

Table 5.22: The CRITICAL VALUES for test statistic $Q_I^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.25, 0.25, 0.25, 0.25)^T$ (no change), and KNOWN "change point". The column $\chi^2_3(1-\alpha)$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$q_{1-\alpha}$	perm1000	perm10000	boot1000	boot10000
30	15	3.056	2.653	2.602	2.441	2.576
50	25	3.056	2.664	2.706	2.679	2.644
100	50	3.056	2.892	2.849	2.829	2.836

Table 5.23: The CRITICAL VALUES for test statistic $Q_{II}^{(K)}$(and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.25, 0.25, 0.25, 0.25)^T$ (no change), and UNKNOWN "change point". The column $q_{1-\alpha}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$\chi^2_3(1-\alpha)$	perm1000	perm10000	boot1000	boot10000
30	15	7.815	8.337	7.800	7.269	7.311
	20	7.815	7.377	7.693	7.671	7.706
	23	7.815	8.118	8.118	7.301	7.589
50	25	7.815	7.903	7.642	7.715	7.682
	34	7.815	7.893	7.950	8.167	7.887
	38	7.815	7.267	7.724	7.852	7.705
100	50	7.815	8.102	7.742	8.027	7.743
	67	7.815	7.927	7.956	7.323	7.797
	75	7.815	7.948	7.688	7.891	7.778
500	250	7.815	7.768	7.962	7.706	7.901
	334	7.815	7.831	7.970	8.191	7.895
	375	7.815	7.235	7.885	7.511	7.719

Table 5.24: The CRITICAL VALUES for test statistic $Q_I^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ and KNOWN change point. The column $\chi^2_3(1-\alpha)$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$q_{1-\alpha}$	perm1000	perm10000	boot1000	boot10000
30	15	3.056	2.633	2.585	2.533	2.471
	20	3.056	2.545	2.616	2.597	2.564
	23	3.056	2.441	2.488	2.438	2.452
50	25	3.056	2.719	2.669	2.714	2.707
	34	3.056	2.595	2.705	2.875	2.711
	38	3.056	2.636	2.742	2.705	2.749
100	50	3.056	2.943	2.827	2.834	2.787
	67	3.056	2.896	2.856	2.759	2.851
	75	3.056	2.873	2.821	2.823	2.872

Table 5.25: The CRITICAL VALUES for test statistic $Q_{II}^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ and UNKNOWN change point. The column $q_{1-\alpha}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$\chi^2_3(1-\alpha)$	perm1000	perm10000	boot1000	boot10000
30	15	7.815	7.856	7.749	7.275	7.586
	20	7.815	7.250	7.625	8.176	7.436
	23	7.815	6.977	7.549	7.843	7.655
50	25	7.815	7.389	7.810	8.273	7.632
	34	7.815	7.807	7.561	7.458	7.747
	38	7.815	7.567	7.781	7.745	7.828
100	50	7.815	7.542	7.859	7.621	7.933
	67	7.815	7.999	7.735	8.305	7.747
	75	7.815	7.781	7.582	8.043	7.724
500	250	7.815	7.820	7.759	8.727	7.779
	334	7.815	7.987	7.792	7.772	7.754
	375	7.815	7.886	8.067	8.022	7.788

Table 5.26: The CRITICAL VALUES for test statistic $Q_I^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ and KNOWN change point. The column $\chi^2_3(1-\alpha)$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$q_{1-\alpha}$	perm1000	perm10000	boot1000	boot10000
30	15	3.056	2.621	2.571	2.578	2.524
	20	3.056	2.603	2.617	2.600	2.564
	23	3.056	2.537	2.593	2.497	2.567
50	25	3.056	2.603	2.769	2.888	2.670
	34	3.056	2.574	2.691	2.651	2.760
	38	3.056	2.785	2.751	2.812	2.746
100	50	3.056	2.871	2.884	2.828	2.848
	67	3.056	2.843	2.863	2.978	2.816
	75	3.056	2.788	2.825	2.856	2.869

Table 5.27: The CRITICAL VALUES for test statistic $Q_{II}^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ and UNKNOWN change point. The column $q_{1-\alpha}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$\chi^2_3(1-\alpha)$	perm1000	perm10000	boot1000	boot10000
30	15	7.815	7.067	7.567	7.441	7.266
	20	7.815	7.465	7.465	7.571	7.500
	23	7.815	7.389	7.301	7.487	7.560
50	25	7.815	7.462	7.693	7.779	7.642
	34	7.815	7.608	7.729	8.153	7.673
	38	7.815	7.225	7.616	7.685	7.565
100	50	7.815	7.346	7.796	7.125	7.722
	67	7.815	7.362	7.889	7.540	7.803
	75	7.815	7.620	7.731	7.538	7.791
500	250	7.815	7.694	7.791	7.945	7.949
	334	7.815	7.618	7.936	7.991	7.939
	375	7.815	7.865	7.707	8.129	7.905

Table 5.28: The CRITICAL VALUES for test statistic $Q_I^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ and KNOWN change point. The column $\chi^2_3(1-\alpha)$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	$q_{1-\alpha}$	perm1000	perm10000	boot1000	boot10000
30	15	3.056	2.400	2.533	2.486	2.463
	20	3.056	2.511	2.615	2.463	2.538
	23	3.056	2.485	2.621	2.538	2.577
50	25	3.056	2.566	2.688	2.762	2.663
	34	3.056	2.725	2.727	2.635	2.682
	38	3.056	2.716	2.720	2.711	2.713
100	50	3.056	2.760	2.834	2.686	2.792
	67	3.056	2.705	2.815	2.826	2.868
	75	3.056	2.853	2.859	2.651	2.865

Table 5.29: The CRITICAL VALUES for test statistic $Q_{II}^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ and UNKNOWN change point. The column $q_{1-\alpha}$ contains the critical value obtained from asymptotic test, in the columns perm1000, perm10000, boot1000 and boot10000 one can find the critical values from permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.103	0.131	0.162	0.159
	20	0.125	0.113	0.113	0.113
	23	0.065	0.065	0.090	0.081
50	25	0.200	0.217	0.211	0.212
	34	0.148	0.146	0.137	0.148
	38	0.171	0.146	0.137	0.148
100	50	0.409	0.437	0.417	0.437
	67	0.339	0.336	0.405	0.349
	75	0.281	0.296	0.285	0.287
500	250	0.994	0.994	0.994	0.994
	334	0.982	0.981	0.979	0.981
	375	0.973	0.963	0.971	0.966

Table 5.30: The TEST POWER for test statistic $Q_I^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ and KNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.114	0.121	0.135	0.150
	20	0.099	0.086	0.089	0.093
	23	0.099	0.092	0.102	0.096
50	25	0.182	0.195	0.183	0.184
	34	0.140	0.117	0.093	0.113
	38	0.101	0.087	0.088	0.087
100	50	0.327	0.355	0.352	0.364
	67	0.236	0.254	0.272	0.254
	75	0.158	0.168	0.168	0.158

Table 5.31: The TEST POWER for test statistic $Q_{II}^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.4, 0.25, 0.25)^T$ and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.138	0.145	0.164	0.149
	20	0.140	0.125	0.097	0.134
	23	0.128	0.096	0.085	0.090
50	25	0.260	0.233	0.197	0.244
	34	0.171	0.181	0.197	0.173
	38	0.156	0.142	0.145	0.139
100	50	0.480	0.454	0.474	0.447
	67	0.387	0.409	0.364	0.409
	75	0.341	0.366	0.319	0.347
500	250	0.997	0.997	0.995	0.997
	334	0.993	0.993	0.993	0.993
	375	0.980	0.979	0.979	0.980

Table 5.32: The TEST POWER for test statistic $Q_I^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ and KNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.131	0.136	0.136	0.145
	20	0.097	0.095	0.098	0.105
	23	0.083	0.080	0.085	0.082
50	25	0.234	0.174	0.150	0.214
	34	0.157	0.130	0.135	0.123
	38	0.086	0.090	0.081	0.091
100	50	0.376	0.369	0.393	0.387
	67	0.274	0.270	0.249	0.281
	75	0.186	0.180	0.173	0.169

Table 5.33: The TEST POWER for test statistic $Q_{II}^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.2, 0.3, 0.4)^T$ and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.621	0.572	0.580	0.592
	20	0.505	0.505	0.494	0.504
	23	0.390	0.400	0.381	0.381
50	25	0.822	0.817	0.816	0.818
	34	0.727	0.719	0.689	0.723
	38	0.665	0.646	0.643	0.651
100	50	0.988	0.983	0.989	0.983
	67	0.970	0.962	0.967	0.965
	75	0.945	0.939	0.945	0.937
500	250	1.000	1.000	1.000	1.000
	334	1.000	1.000	1.000	1.000
	375	1.000	1.000	1.000	1.000

Table 5.34: The TEST POWER for test statistic $Q_I^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ and KNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

N	K	perm1000	perm10000	boot1000	boot10000
30	15	0.546	0.489	0.507	0.522
	20	0.394	0.350	0.415	0.382
	23	0.219	0.182	0.196	0.188
50	25	0.782	0.750	0.728	0.754
	34	0.560	0.560	0.593	0.579
	38	0.364	0.364	0.364	0.364
100	50	0.970	0.965	0.972	0.969
	67	0.937	0.923	0.923	0.916
	75	0.787	0.786	0.834	0.784

Table 5.35: The TEST POWER for test statistic $Q_{II}^{(K)}$ and true values $\boldsymbol{p} = (0.25, 0.25, 0.25, 0.25)^T$ and $\boldsymbol{q} = (0.1, 0.1, 0.1, 0.7)^T$ and UNKNOWN change point. The columns perm1000, perm10000, boot1000 and boot10000 contain the values for permutation tests and bootstrap with 1 000 and 10 000 repetitions.

Conclusion

Three special situations of change-point problem from testing perspective were studied in this thesis. Both known and unknown cases were considered and using limit theorems as well as approximation using bootstrap and permutation test were described and compared using simulations. The broken-line case with few different sets of conditions on the random errors were described in Chapter 2 and the different approaches were compared on simulations in Chapter 4. The possible approaches for the discrete cases of Bernoulli and Multinomial distributions were introduced in Chapter 3 and compared on simulations in Chapter 5.

The results of the simulation study across all the studied cases showed that in case of known change point, all the introduced approaches are comparable already for quite small sample sizes. In case of unknown change point, the asymptotic methods using approximation by limit theorems and extreme value theory seem to converge slowly and so they are quite conservative in case of small sample sizes. On the other hand, the bootstrap and permutation tests seem to work better for the small sample sizes, but they are computationally intensive and time-consuming for larger sample sizes ($N \geq 500$).

Bibliography

J. Anděl. *Základy matematické statistiky.* Matfyzpress, Praha, 2005. ISBN 80-86732-40-1.

J. Antoch and M. Hušková. Permutation tests in change point analysis. *Statistics & Probability Letters*, 53(No. 1):37–46, 2001.

J. Antoch, M. Hušková, and D. Jarušková. Change-point analysis. In *Lauro, N. C. et al. : Lecture Notes of the Fifth IASC Summer School*, pages 1–75, Voorburg, 2000. ISI.

J. Antoch, G. Gregoire, and D. Jarušková. Detection of structural changes in generalized linear models. *Statistics & Probability Letters*, 69(No. 3):315–332, 2004.

P. Billingsley. *Convergence of Probability Measures.* Wiley, New York, 1968. ISBN 0-471-07242-7.

J. Chen and A. K. Gupta. *Parametric Statistical Change Point Analysis.* Birkh auser, Boston, 2000. ISBN 0-8176-4169-6.

G. Ciuperca. Maximum likelihood estimator in a two-phase nonlinear random regression model. *Statistics & Decisions*, 22(No. 4):335–349, 2004.

G. Ciuperca. The m-estimation in a multi-phase random nonlinear model. *Statistics & Probability Letters*, 79(No. 5):573–580, 2009.

M. Csörgő and L. Horváth. *Limit Theorems in Change-Point Analysis.* Wiley, New York, 1997. ISBN 0-471-95522-1.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge, 1997. ISBN 978-0-521-57471-6.

J.-F. Dupuy. Estimation in a change-point hazard regression model. *Statistics & Probability Letters*, 76(No. 2):182–190, 2006.

B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(No. 1):1–26, 1979.

P. I. Good. *Permutation, parametric and bootstrap tests of hypotheses.* Springer Science+ Business Media, 2005.

M. Hušková. *Pravděpodobnost a matematická statistika.* Karolinum, Praha, 2005. ISBN 80-246-0009-9.

M. Hušková and J. Antoch. Detection of structural changes in regression. *Tatra Mountains Publications*, 26, 2003.

D. Jarušková. Testing appearance of linear trend. *Journal of Statistical Planning and Inference*, 70(No. 2):263–276, 1998.

S. A. Julious. Inference and estimation in a changepoint regression problem. *The Statistician*, 50(No. 1):51–61, 2001.

I. López, M. Gámez, J. Garay, T. Standovár, and Z. Varga. Application of change-point problem to the detection of plant patches. *Acta Biotheor*, 58(No. 1):51–63, 2010.

K. Pastorková. *Použití metody bootstrap pro intervalový odhad parametru polohy [Master thesis].* MFF UK, Prague, 1994.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society Supplement*, 4(No. 1): 119–130, 1937a.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Journal of the Royal Statistical Society Supplement*, 4(No. 2):225–232, 1937b.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations: Iii, the analysis of variance test. *Biometrika*, 29(No. 3/4):322–335, 1938.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

C. R. Rao. *Linear Statistical Inference and Applications.* Wiley, New York, 1965. ISBN 0-471-21875-8.

P. K. Sen. *Sequential nonparametrics: Invariance Principles and Statistical Inference.* Wiley, New York, 1981. ISBN 0-471-06013-5.

H. Zhou and K.-Y. Liang. On estimating the change point in generalized linear models. In *Beyond Parametrics in Interdisciplinary Research, IMS Collections. Vol. 1.*, pages 305–320. IMS USA, 2008.

List of Tables

Appendix A

Used theorems

Theorem A.1 (Ljapunov Central Limit Theorem). *Let $X_{n,1}, \ldots, X_{n,k_n}$ be independent random variables for $n \geq 1$. Denote $\mathsf{E}X_{n,i} = \mu_{n,i}$, $\operatorname{var} X_{n,i} = \sigma^2_{n,i}$ and $\mathsf{E}|X_{n,i} - \mathsf{E}X_{n,i}|^3 = \rho^3_{n,i}$, $i = 1, \ldots, k_n$. Assume*

$$\lim_{n\to\infty} \frac{\left(\sum_{i=1}^{k_n} \rho^3_{n,i}\right)^{1/3}}{\left(\sum_{i=1}^{k_n} \sigma^2_{n,i}\right)^{1/2}} = 0,$$

then

$$\lim_{n\to\infty} \mathsf{P}\left(\frac{\sum_{i=1}^{k_n}(X_{n,i} - \mu_{n,i})}{\sqrt{\sum_{i=1}^{k_n} \sigma^2_{n,i}}} \leq x\right) = \Phi(x),$$

where Φ is a cumulative distribution function of $\mathsf{N}(0,1)$.

Proof. Can be found in Hušková [2005], pages 85–87. □

Theorem A.2 (Multivariate Cramér-Slutsky Theorem). *Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$, $(\boldsymbol{X}_2, \boldsymbol{Y}_2), \ldots$ be a sequence of pairs of random vectors and $\boldsymbol{C}$ be a vector of real numbers. If $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{C}$ for $n \to \infty$, then*

$$\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{X} + \boldsymbol{C}, \quad \text{for } N \to \infty.$$

Let $\boldsymbol{B}_1, \boldsymbol{B}_2, \ldots$ be a sequence of random matrices and $\boldsymbol{B}$ be a matrix of real numbers. If $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ and $\boldsymbol{B}_n \xrightarrow{\mathsf{P}} \boldsymbol{B}$ for $n \to \infty$, then

$$\boldsymbol{B}_n \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{B}\boldsymbol{X}, \quad \text{for } N \to \infty.$$

Proof. Can be found in Rao [1965], page 102. □

Theorem A.3 (Donsker's theorem). *Let the random variables $X_1, X_2, \ldots$ defined on $(\Omega, \mathcal{B}, \mathsf{P})$ be independent identically distributed with mean 0 and finite positive variance σ^2 and let $S_n = X_1 + \cdots + X_n$ be the partial sums. Define a random element Y_n by*

$$Y_n(t) = \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor} + (nt - \lfloor nt \rfloor) \frac{1}{\sigma\sqrt{n}} X_{\lfloor nt \rfloor + 1},$$

where $\lfloor nt \rfloor$ *is the greatest integer smaller than* nt*. Then*

$$Y_n \xrightarrow{\mathsf{d}} W,$$

where W *is a Wiener process.*

Proof. Can be found in Billingsley [1968], pages 68–70. $\square$

Theorem A.4. *Let the random variables* $X, X_1, X_2, \ldots$ *be defined on* $(\Omega, \mathcal{B}, \mathsf{P})$ *and* $X_n \xrightarrow{d} X$ *for* $n \to \infty$*. Let* h *be a measurable mapping of* Ω *into another metric space* Ω^* *and let* D_h *be the set of discontinuities of* h*. If* $\mathsf{P}(X \in D_h) = 0$ *then*

$$h(X_n) \xrightarrow{d} h(X), \quad \textit{for } n \to \infty.$$

Proof. Can be found in Billingsley [1968], pages 29–31. $\square$

Corollary A.5. *Let* $\{X_n(t), t \in (0,1)\}$ *be a random process such that*

$$\{X_n(t), t \in (0,1)\} \xrightarrow{d} \{W(t), t \in (0,1)\}, \quad \textit{for } n \to \infty,$$

where $\{W(t), t \in (0,1)\}$ *is a Wiener process. Denote* $Y_n(t) = X_n(t) - t\,X_n(1)$ *for* $t \in (0,1)$*, then*

$$\{Y_n(t), t \in (0,1)\} \xrightarrow{d} \{B(t), t \in (0,1)\}, \quad \textit{for } n \to \infty,$$

where $\{B(t), t \in (0,1)\}$ *is a Brownian bridge.*

Theorem A.6. *Let the vector* $\boldsymbol{X}$ *has multivariate normal distribution with expectation* $\boldsymbol{\mu}$ *and variance matrix* $\boldsymbol{V}$ *with rank* $r \geq 1$*. Then for any pseudo-inverse matrix* $\boldsymbol{V}^-$ *the random variable* $(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{V}^- (\boldsymbol{X} - \boldsymbol{\mu})$ *has* χ^2_r *distribution.*

Proof. Can be found in Anděl [2005], page 68. $\square$

Theorem A.7 (Multivariate Donsker's theorem)**.** *Let the* p*-dimensional random vectors* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *be independent identically distributed with mean* $\boldsymbol{0}$ *and finite positive definite variance matrix* $\boldsymbol{V}$*. Define a random element* Y_n *by*

$$Y_n(t) = \frac{1}{n} \left[\sum_{i=1}^{\lfloor nt \rfloor} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_n \right) \right]^T \boldsymbol{V}^{-1} \left[\sum_{i=1}^{\lfloor nt \rfloor} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_n \right) \right],$$

where $\lfloor nt \rfloor$ *is the greatest integer smaller than* nt*. Then*

$$\{Y_n(t),\, t \in (0,1)\} \xrightarrow{d} \left\{ \sum_{j=1}^{p} B_j^2,\, t \in (0,1) \right\},$$

where B_j *for* $j = 1, \ldots, p$ *are independent Brownian bridges.*

Proof. Can be found in Sen [1981]. $\square$

Appendix B

Important parts of the code

The most important parts of the code can be found in this Chapter. All codes and simulated data sets can be found on the attached CD.

B.1 Broken-line model

The simulated data sets:

```
####################################################b2=-1
###N=30,m=15, bldata30m2b1.txt
rm(list=ls())
N<-30
m<-ceiling(N/2)
mre<-0
sre<-sqrt(1)
b0<-20
b1<-5
b2<--1

set.seed(909634)
err<-rnorm(N,mre,sre)
x<-numeric(N)
y<-numeric(N)
for (i in 1:N){
x[i]<-i/N
if (i<=m){y[i]<-b0+b1*x[i]+err[i]}
else {y[i]<-b0+b1*x[i]+b2*(x[i]-x[m])+err[i]}
}
data<-cbind(x,y)
write.table(data,"bldata30m2b1.txt",row.names=F)
```

Test statistic $R_I^{(K)}$:

```
RI<-function(data,N,m){
x<-data$x
```

```
y<-data$y
mlt<-(x>m/N)
xm<-x[m]
xavg<-sum(x)/N
yavg<-sum(y)/N

a<-sum(mlt*(y-yavg)*(x-xm))
b<-sum(mlt*(x-xavg)*(x-xm))
c<-sum((x-xavg)^2)
d<-sum((y-yavg)*(x-xavg))
e<-sum(mlt*(x-xm)^2)
f<-sum(mlt*(x-xm))

jm<-e-f^2/N-b^2/c
beta2hat<-(a-b/c*d)/jm
varbeta2hat<-(e-f^2/N-b^2/c)/jm^2

RI<-abs(beta2hat)/sqrt(varbeta2hat)
}
```

Test statistic $R_{II}^{(K)}$:

```
RII<-function(data,N,m){
x<-data$x
y<-data$y
mlt<-(x>m/N)
xm<-x[m]
xavg<-sum(x)/N
yavg<-sum(y)/N

a<-sum(mlt*(y-yavg)*(x-xm))
b<-sum(mlt*(x-xavg)*(x-xm))
c<-sum((x-xavg)^2)
d<-sum((y-yavg)*(x-xavg))
e<-sum(mlt*(x-xm)^2)
f<-sum(mlt*(x-xm))

jm<-e-f^2/N-b^2/c
beta2hat<-(a-b/c*d)/jm
beta1hat<-d/c-beta2hat*b/c
beta0hat<-yavg-beta1hat*xavg-beta2hat*f/N
sigma2hat<-sum((y-beta0hat-beta1hat*x-beta2hat*(x-xm)*mlt)^2)
            /(N-3)
varbeta2hat<-sigma2hat*(e-f^2/N-b^2/c)/jm^2

RII<-abs(beta2hat)/sqrt(varbeta2hat)
}
```

Test statistic R_{III}:

```
RII<-function(data,N){
x<-data$x
y<-data$y
RI<-numeric(N-3)
for (m in 2:(N-2)){
mlt<-(x>m/N)
xm<-x[m]
xavg<-sum(x)/N
yavg<-sum(y)/N

a<-sum(mlt*(y-yavg)*(x-xm))
b<-sum(mlt*(x-xavg)*(x-xm))
c<-sum((x-xavg)^2)
d<-sum((y-yavg)*(x-xavg))
e<-sum(mlt*(x-xm)^2)
f<-sum(mlt*(x-xm))

jm<-e-f^2/N-b^2/c
beta2hat<-(a-b/c*d)/jm
varbeta2hat<-(e-f^2/N-b^2/c)/jm^2
RI[m-1]<-abs(beta2hat)/sqrt(varbeta2hat)
}
RII<-max(RI)
}
```

B.2 Discrete case

B.2.1 Bernoulli case

The simulated data sets:

```
##########################################p1=0.6
###N=30,m=15 data30m2p60.txt
N<-30
m<-ceiling(N/2)
p0<-0.5
p1<-0.6

data<-numeric(N)
set.seed(1858475)
for (i in 1:m){
  data[i]<-rbinom(1,1,p0)
}
for (i in (m+1):N){
  data[i]<-rbinom(1,1,p1)
}
```

```
write.table(data,"data30m2p60.txt",row.names=F)
```

Test statistic $T_I^{(K)}$:

```
T<-function(data,N,m){
  s1<-sum(data[1:m])
  s2<-sum(data[(m+1):N])
  s0<-sum(data)

  if (s0>0){
if (s0<N) {cit<-abs(s1/m-s2/(N-m))
jm<-sqrt(N/(m*(N-m))*s0/N*(1-s0/N))
T<-cit/jm}
else {T<-0}}
  else {T<-0}
}
```

Test statistic $T_{II}^{(K)}$:

```
T<-function(data,N,m){
  s1<-sum(data[1:m])
  s2<-sum(data[(m+1):N])
  s0<-sum(data)
  if (s0>0){
if (s0<N) {cit<-abs(s1/m-s2/(N-m))
   jm<-sqrt((s1/(m^2)*(1-s1/m))+(s2/((N-m)^2)*(1-s2/(N-m))))
   T<-cit/jm}
else {T<-0}}
  else {T<-0}
}
```

Test statistic T_{III}:

```
T3<-function(data,N,m){
  s1<-sum(data[1:m])
  s2<-sum(data[(m+1):N])
  s0<-sum(data)
  if (s0>0){
if (s0<N) {cit<-abs(s1-m*s0/N)
jm<-sqrt(N)*sqrt(s0/N*(1-s0/N))
#   jm<-sqrt(N)*sqrt((s1/(m)*(1-s1/m))+(s2/((N-m))*(1-s2/(N-m))))
   T3<-cit/jm}
else {T3<-0}}
  else {T3<-0}
}
```

```
T3vec<-numeric(N-1)
for (i in 1:(N-1)){T3vec[i]<-T3(data,N,i)}

T3max<-max(T3vec)
```

B.2.2 Multinomial case

The simulated data sets:

```
##########################################p=1
###N=30,m=15 mdata30m2p1.txt
N<-30
m<-ceiling(N/2)
l<-3
val<-c(1,2,3,4)
p0<-c(0.25,0.25,0.25,0.25)
p1<-c(0.1,0.4,0.25,0.25)

data<-numeric(N)
set.seed(70547965)
for (i in 1:m){
  data[i]<-sample(val,1,prob=p0)
}
for (i in (m+1):N){
  data[i]<-sample(val,1,prob=p1)
}

write.table(data,"mdata30m2p1.txt",row.names=F)
```

Test statistic $Q_I^{(K)}$:

```
Q<-function(data,N,K,s){
  v0<-numeric(s)
  v1<-numeric(s)
  vs<-numeric(s)
  for (i in 1:K){
    j<-data[i]
    v0[j]<-v0[j]+1
    vs[j]<-vs[j]+1
  }
  for (i in (K+1):N){
    j<-data[i]
    v1[j]<-v1[j]+1
    vs[j]<-vs[j]+1
  }
 pom<-numeric(s)
 for (j in 1:(s)){if (vs[j]>0){pom[j]<-((v0[j]-K*vs[j]/N)
```

```
                    /sqrt(vs[j]/N))^2}
else pom[j]<-0}
  Q<-N/(K*(N-K))*sum(pom)
}
```

Test statistic $Q_I^{(K)}$:

```
Q<-function(data,N,s){
 QK<-numeric(N-1)
 for (K in 1:(N-1)){
  v0<-numeric(s)
  v1<-numeric(s)
  vs<-numeric(s)
  for (i in 1:K){
    j<-data[i]
    v0[j]<-v0[j]+1
    vs[j]<-vs[j]+1
  }
  for (i in (K+1):N){
    j<-data[i]
    v1[j]<-v1[j]+1
    vs[j]<-vs[j]+1
  }
 pom<-numeric(s)
 for (j in 1:(s)){if (vs[j]>0){pom[j]<-((v0[j]-K*vs[j]/N)
                    /sqrt(vs[j]/N))^2}
else pom[j]<-0}
  QK[K]<-1/N*sum(pom)
 }
Q<-max(QK)
}
```